A CORPUS-BASED STUDY OF THE USE OF "*BE*" IN MALAY ESL LEARNER ESSAYS

ROSLINA BINTI ABDUL AZIZ

FACULTY OF LANGUAGES AND LINGUISTICS
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

A CORPUS-BASED STUDY OF THE USE OF "*BE*" IN MALAY ESL LEARNER ESSAYS

ROSLINA BINTI ABDUL AZIZ

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF LANGUAGES AND LINGUISTICS
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITI MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Roslina binti Abdul Aziz

Registration/Matric No: THA090016

Name of Degree: Doctor of Philosophy

Title of Thesis ("this work"):

A CORPUS-BASED STUDY ON THE USE OF "*BE"* IN MALAY ESL LEARNER ESSAYS

Field of Study: Corpus Linguistics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first has and obtained;
(6) I am fully aware that if the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's signature Date

Subscribed and solemnly declared before,

Witness's signature Date
Name: Prof. Madya Dr. Su'ad binti Awab
Designation: Supervisor

A CORPUS-BASED STUDY ON THE USE OF "*BE*" IN MALAY ESL LEARNER ESSAYS

ABSTRACT

This corpus-based study investigates (i) the distributional patterns for each form and function of *BE* in the Malay ESL learner essays, (ii) the patterns of grammatical and ungrammatical uses of *BE* by the learners and (iii) the extent of influence of the syntactic environments on the grammatical and ungrammatical uses of *BE*. The data for the study were harvested from the Malaysian Corpus of Learner English (MACLE). A total of 366 Malay learner essays were extracted from MACLE to form the Malay ESL learner sub-corpus. LOCNESS (Louvain Corpus of Native English Essays), which contains essays written by native speaker learners, was chosen to be the control corpus. In preparing the corpora for analysis, LOCNESS was tagged using CLAWS POS tagger, while MACLE was manually tagged using tagsets developed based on the analytical parameters set for this study. WordSmith Tools Version 5 was utilised to analyse the corpora. The study employed both quantitative and qualitative analyses. The quantitative analysis encompassed the frequency counts of all grammatical and ungrammatical uses of *BE*, while the qualitative analysis involved textual analysis on the grammatical and ungrammatical *BE* constructions. The quantitative findings reveal significantly higher grammatical use of *BE* in the Malay learner data, which is supported by the qualitative findings, which reveal the use of a wide range of forms and functions of the verb in predominantly more structurally complex constructions. In addition, the study has also unveiled the patterns of the most persistent ungrammatical use of *BE* namely, overgeneration and omission. Finally, based on the findings, the study proposes a corpus consultation model for the teaching of *BE* to ESL learners in Malaysian universities.

Keywords: *BE*, corpus-based, learner corpus, computer aided error analysis

ABSTRAK

Kajian berasaskan korpus ini mengkaji (i) corak agihan setiap satu kata kerja *BE* berserta fungsinya dalam karangan yang dihasilkan oleh pelajar ESL Melayu, (ii) corak pengunaan *BE* yang mengikuti tatabahasa dan yang menyalahi tatabahasa oleh pelajar, dan (iii) sejauh mana persekitaran sintaksi mempengaruhi penggunaan *BE*. Data bagi kajian ini diperolehi dari Malaysian Corpus of Learner English (MACLE). Sejumlah 366 karangan yang ditulis oleh pelajar Melayu diekstrak dari MACLE bagi membentuk sub-korpus pelajar ESL Melayu. LOCNESS (Louvain Corpus of Native English Essays) yang mengandungi karangan yang ditulis oleh penutur asal Bahasa Inggeris telah dipilih sebagai korpus kawalan. Bagi persediaan analisis data korpus, LOCNESS telah dikod menggukan applikasi CLAWS, manakala MACLE telah dikod secara manual menggunakan kod-kod yang telah dibangunkan mengikut parameter analitik yang telah ditetapkan dalam kajian ini. WordSmith Tools Versi 5 telah digunakan bagi analisa data korpus. Kajian ini menggunakan analisis kuantitatif dan kualitatif. Analisa kuantitatif merangkumi bilangan frekuensi semua penggunaan *BE*, sementara itu analisa kualitatif merangkumi analisa teks keatas semua penggunaan *BE*. Dapatan dari analisa kauntitatif mendedahkan penggunaan *BE* mengikut tatabahasa adalah lebih tinggi berbanding yang menyalahi tatabahasa. Ini disokong oleh dapatan dari analisa kualitatif yang mendedahkan penggunaan pelbagai bentuk dan fungsi *BE* dalam binaan ayat-ayat kompleks. Selain itu, kajian ini mendedahkan corak kesalahan dalam pengunaan *BE* yang paling ketara iaitu; penambahan dan pengguguran. Akhir sekali, berdasarkan dapatan yang diperolehi, kajian ini mencadangkan model konsultasi korpus bagi pengajaran *BE* kepada pelajar ESL di universiti-universiti di Malaysia.

ACKNOWLEDGEMENTS

First and foremost praise Allah the Almighty who has granted me the strength to complete my PhD journey.

I would like to take this opportunity to convey my deepest gratitude to my supervisors, Professor Dr. Zuraidah Mohd Don and Prof. Madya Dr. Su'ad Awab for their continuous support, guidance and encouragement throughout the research. At many stages in course of this research I have benefitted from their vast knowledge and expertise in research and research writing.

Special thanks to the Ministry of Higher Education Malaysia for granting me the scholarship, without which this study would not have been possible. I would also like to thank my employer University Teknologi MARA, the Rector of Universiti Teknologi MARA Cawangan Pahang, the Dean of Academy of Language Studies UiTM Malaysia for the opportunity.

Most importantly, my deepest and heartfelt gratitude goes to my family members especially to my caring and supportive husband, Mohd Rozaidi Ismail, who has been my pillar of strength. Thank you for your patience, understanding and continuous support, which have carried me through the toughest times throughout this endeavor. To my three angels Sarah Syazwina, Julia Syazwani and Alya Syazwana, who have given me so much happiness and joy. I hope I have made you proud. I am also indebted to my parents Abdul Aziz and Meriam and sisters Roslayati, Rosmiza and Fazilah, whose love and emotional support have given me the courage and strength to complete this study. Last but not least, a special thank you goes to my close friends, for their encouragement and stimulating words that have helped lifted my spirits.

TABLE OF CONTENTS

CHAPTER 8: CORPUS CONSULTATION MODEL

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.0 Background of the Study

1.0.1 Overview of *BE*

This study intends to analyse in detail the use of *BE* by L1-Malay ESL learners involved in the study. The interest was sparked by the researcher's own experience as an English teacher, encountering first-hand the ill-formed use of the verb in the writings of ESL learners she taught. The followings are sample sentences containing some of the most common errors in the use of *BE* the researcher has encountered in her years of teaching:

1) a. *Some countries in the world ***are compete*** each other to achieve same level or to be the highest level of International education.
   b. *Some people ***are agree*** with this policy.
   c. *...our late grandfather ***was try*** until their last drop of blood to ensure the independent of this country.
   d. *We can communicate with the friends that we ***are already know*** and we also can find the new friends.
   e. *They Ø willing to use their money to top-up their handphone *(to reload mobile phone air time).*
   f. *The computer Ø very import(*ant)* to the all people.

The sentences above highlight two major types of errors in the use of *BE*, which include *BE* being inserted (a, b, c, d) and omitted in obligatory context (e, f). It is intriguing to find students repeatedly producing such ill-formed constructions, despite having been taught the use of *BE* since the 1st year of school and that *BE* is a very common verb in the English language (Biber, Johansson, Conrad, Leech, & Finegan, 1999). According to the findings of first language (L1) acquisition research *BE* is probably the easiest

word to learn. Dulay, Burt and Krashen (1982) listed *BE* among the first words acquired by children acquiring English as their L1. Copula *BE* and auxiliary *BE* were also reported to be among the earlier grammatical morphemes acquired by L2 child learners of English (Dulay et al., 1982). Dulay et al. (1982) posited a "natural order in which L2 learners acquire certain syntactic and morphological structures" (p. 204), and that copula *BE* and auxiliary *BE* (i.e. *'s*, *is*) belong to the second group of items acquired after word order and case (p. 208). The researchers believed that regardless of the learners' L2 backgrounds, they would acquire the grammatical morphemes in a similar order.

Unfortunately, many ESL learners in Malaysia still find *BE* confusing and they are unsure of when, where and how to use it. Research conducted on the written and spoken language of ESL learners in Malaysia found *BE* being dropped in obligatory context (e.g. *It ∅ also good for-for our reading*), inserted before a main verb (e.g. *The accident **was happened** at Jalan Raya Laut*), and used in the wrong tense (e.g. *In a kingdom, there **is** a beautiful princess*) and agreement (e.g. *The lecturer **are**..*) (Maros, Tan & Khazriyati, 2007; Siti Hamim & Mohd Mustafa, 2010; Ting, Mahanita & Chang, 2010; Wee, 2009; Wee, Sim & Kamaruzaman, 2010). The same types of misuses are also found among learners from different L1 backgrounds such as Russian (Ionin & Wexler, 2001; Unlu & Hatipoglu, 2012), Chinese (Chan, 2004; Ju, 2000; Lee & Huang, 2004; Yip, 1994), Spanish (Fleta, 2003; Oshita, 2000), Italian, Japanese, Korean (Oshita, 2000), Sinhala (Herat, 2005), Arabic (Muneera & Wong, 2011; Murad & Khalil, 2015) and Nigerian (Akande, 2013).

It is puzzling why *BE* is terribly difficult and confusing, despite its status as the most common verb in English (Biber et al., 1999) and probably the easiest to acquire (Dulay et al., 1982). The inflectional variations and irregularities of *BE* are believed to be a major cause of confusion according to Celce-Murcia and Larsen-Freeman (2016), Wee,

Sim and Kamaruzam (2010) and Unlu and Hatipoglu (2012). Unlike other primary verbs, *BE* is the only verb in the English grammatical system with a total of eight (8) inflections; ***am, is, are, was, were, be, being, been***. Each inflection has to be aligned with person, number and tense. In addition, it also has the unique form ***am*** to agree with ***I*** and two past tense forms; ***was and were,*** following singular and plural subjects. In order to successfully master the use of *BE* a learner must possess the knowledge of tense, number and person.

Another reason why *BE* poses serious problems for ESL learners is probably because learners tend to be confused by its multiple functions. *BE* could either function as a copular, mainly as a link between the subject to its complement or as an auxiliary to mark progressive aspect or passive voice. Additionally, it also serves as a negative or an interrogative operator. It also has other special usages for example, when combined with *there* to form existential '*there BE'* and with *it* in the '*it*-cleft'. Considering the inflectional variations and irregularities of the forms and the multiple functions of the verb, it is no wonder L2 learners in general find it a challenge to master the correct use of the verb.

Past studies also highlight the possible influence of the syntactic environments on the supply of *BE*, which include firstly the types of predicates complementing *BE* (Gavruseva & Maisterheim, 2003; Herat, 2005; Lee & Huang, 2004; Platt & Weber, 1980). Gavruseva and Maisterheim (2003) for instance reported that *BE* is more regularly supplied before individual-level predicates (predicates that denote permanent properties- *she's happy*) than before stage-level predicates (predicates that denote temporary properties- *it's in the kitchen*). Herat (2005) reported that Sri Lankan speakers of English tended to omit copula *BE* more frequently before adjective predicates. The same *BE* omission pattern was also reported among Chinese learners (Lee & Huang, 2004). Platt and Weber (1980) in their description of Malaysian English

(ME) reported that *BE* was supplied less frequently in pre-locative predicate position than any other positions (nominal and adjectival) in the ME, which also suggests predicate sensitivity to the supply of *BE*.

The supply of *BE* could also be sensitive to the presence of intensifiers and auxiliaries. Chan (2004) highlighted that Chinese ESL learners tended to omit *BE* in the position after modal auxiliaries, while Lee & Huang (2004) reported that *BE* tended to be dropped when it was modified by a degree adverb or negated by *not* (*BE + very/not/so + adjective*). Furthermore, there is also the tendency for *BE* to be inserted before a special category of intransitive verb that is unaccusative (e.g. *happen, sink, fall*). Yip (1994), Ju (2000) and Oshita (2000) reported that learners would insert *BE* before unaccusative verbs to produce passive like constructions such as *What **was happened** yesterday* (Yip, 1994). The same *BE* insertion structure was also reported in the data of L1-Malay learners by Arshad and Hawanum (2010) and Wee (2009). Wee (2009) in her study highlighted that such insertion would normally involve the past tense form *was/were* (*accident **was happened***). According to the researcher *was/were* was interpreted by the learners as the marker for past tense and the insertion was the result of checking the tense feature.

In the context of ESL in Malaysia especially among L1-Malay ESL learners, the variability in the supply of *BE* is often associated with the negative interlingual transfer from the Malay language (Maros, Tan & Khazriyati, 2007; Siti Hamim & Mohd Mustafa, 2010; Ting, Mahanita & Chang, 2010; Wee, 2009; Wee, Sim & Kamaruzaman, 2010). In Malay *BE* is non-existence, the only type of verb close to *BE* is '*Kata Pemeri';* '*ialah'* and '*adalah'* (Nik Safiah, Farid, Hashim, & Abdul Hamid, 2010). Unlike copula *BE*, '*ialah'* and '*adalah'* usages are very restricted. In Malay '*ialah'* is only used in equative sentences, when both the subject and its predicate bare

the same meaning. It is realised in *NP + NP* structure. The function of '*ialah'* in this aspect is almost similar to English copula *BE* as illustrated by the following sentences:

2) a. Malay structure:

Antara perkara yang dibicarakan dalam buku itu ***ialah*** masalah moral negara.
(*NP + NP*)

Nik Safiah et al. (2010, p. 264)

b. English structure:
Among the issues discussed in the book ***are*** moral problems in the country.

The use of '*adalah'* in Malay is to link a subject to a predicate which describes or qualifies the subject in *NP+PP* and *NP+AP* constructions as shown in (3) below:

3) a. Malay structure:

Makanan seimbang ***adalah*** baik untuk kesihan badan
(*NP+AP*)
Sumbangan beliau ***adalah*** dari segi peningkatan ekonom rakyat luar bandar
(*NP+PP)*

Nik Safiah et al. (2010, p. 264)

b. English structure:
Balance meals ***are*** good for health.
His contribution ***was*** to alleviate the economy of the people in the rural areas.

Another condition when *BE* is used in Malay is perhaps with the verb '*ada',* when it is used to show presence of something similar to existential *there* structure in English as in:

4) a. Malay structure:

Ada dua ekor kucing di dalam peti itu.
There are two tail (CLASSIFIER) cats inside box that

Nik Safiah et al. (2010, p. 264)

b. English structure:

There are two cats in the box.

Other than the conditions describe above, *BE* is not essential in the Malay grammar. Researchers working with L1-Malay learner data explain that the non-existence of *BE* in the Malay grammar contributes to the verb being omitted or misused by the L1-Malay learners (Maros, Tan & Khazriyati, 2007; Siti Hamim & Mohd Mustafa, 2010; Ting, Mahanita & Chang, 2010; Wee, 2009; Wee, Sim & Kamaruzaman, 2010).

The literature suggests that *BE* with its multiple forms and functions could pose serious problems to ESL learners worldwide as attested by the misuses of the verb in the data of ESL learners from various L1 backgrounds. There are also empirical evidences showing that covert and overt *BE* could be influenced by the learners' L1 and were sensitive to the syntactic environments like the types of predicates, types of subjects, presence of auxiliaries and intensifiers and types of post-*BE* verbs. The interplay of these factors could have direct and indirect effects on ESL learners' use of *BE* and it is paramount to uncover what learners can do and cannot do with the verb and to what extent these factors influence the correct as well as the incorrect use of the verb so that more effective intervention could be developed to help learners overcome the problems they encounter with *BE*.

1.0.2 Overview of Corpus Linguistics

Corpus linguistic (henceforth CL) according to Leech (1992) is a "new research enterprise, [...] a new philosophical approach to the subject, [...] an 'open sesame' to a new way of thinking about language" (as cited in Granger, Dagneaux & Meunier, 2002, p.4). CL is a comparatively new way of analysing and understanding natural language. The debate of whether it is a methodology or a theory is still ongoing, and linguists have yet to come to a consensus on its definition.

Halliday (1993c) argued that CL is more than just a methodology. According to Halliday CL is able to bring together the activities of data gathering and theorising which have resulted in the "qualitative change in the understanding of language" (as cited in Tognini-Bonelli, 2001, p. 1). It is further argued that the combination of computational tools, algorithmic and statistical methods, and qualitative observations deriving from this approach proves that CL is more than just a methodology; it is "a new research enterprise and philosophical approach to linguistic enquiries" (Tognini-Bonelli, 2001, p.1). Tognini-Bonelli (2001) also stressed that unlike any other discipline under the realm of linguistics, CL possesses what she termed as pre-application methodology, a position where in its application CL will be set not to accept just any given set of rules, but instead identifies its "own set of rules and pieces of knowledge before they are applied". This requires linguists to utilise some new parameters to account for the data, which in turn entails for "a change in what can be referred to as *unit of currency* for linguistic description" (p. 1). Citing areas such as lexicography, translation, stylistics, grammar, language teaching and forensic linguistics, in which CL has made its mark, she highlighted how CL is able to contribute to other linguistic disciplines, thus, lifting its status not to just a powerful methodology but also a theory.

Nevertheless, the view prescribed by Halliday (1993c) and Tognini-Bonelli (2001), is not shared by the majority of linguists working in CL. Granger (2002) for instance, saw it as a "linguistic methodology" which made use of "electronic collections of naturally occurring texts" or corpora (p. 4). According to Granger (2002) "it is neither a new branch of linguistics nor a theory of language but the very nature of the evidence it uses makes it a particularly powerful methodology, one which has the potential to change perspectives on language" (p. 4). The same position was taken much earlier by McEnery and Wilson (2001) in their book *Corpus Linguistics*, which emphasises that CL unlike any other branch of applied linguistics (syntax, semantics, sociolinguistics)

"is a methodology rather than an aspect of language requiring explanation or description" (McEnery & Wilson, 2001, p. 2).

The tradition of corpus compilation and analysis or 'early corpus linguistics', to borrow McEnery and Wilson's (2001) term, can be traced from the various works in the field of linguistics, language acquisition, lexicography, language pedagogy, comparative linguistics and even in syntax and semantics. Early works on child language acquisition, roughly from 1876 to 1926 mostly employed basic corpus-based methodology, such as the work of Preyer (1889) and others after him (Bloom, 1970; Brown, 1973; Stern, 1924) (McEnery & Wilson, 2001). Their studies were based on corpora compilation consisted mostly of parental diaries documenting systematically and in detail a child's utterances. By analysing these carefully compiled data, the researchers were able to generate predictions on child language development. Though not associating themselves with corpus linguistics, the researchers' methodology was evidently corpus-based (McEnery & Wilson, 2001). Other areas employing basic corpus-based methodology, just to name a few include comparative linguistics in Eaton (1940), which involved comparison of Dutch, French, German and Italian frequency of word meanings and descriptive linguistics in Fries (1952) on the descriptive grammar of English (McEnery & Wilson, 2001).

Corpus linguistics tradition did, however, suffer a phase of discontinuity for over 30 years beginning in the late 1950s, when Chomskyan linguistics began to gain momentum and influence. An advocate to rationalism, Chomsky (1957) argued against the use of observable data in determining language grammaticality. According to him, the corpus of utterances collected cannot be considered to represent grammatical sentences since the act violates the very notion of grammaticality; projection, infiniteness and ideal speaker. He added that this collection of data will be so skewed, that they will not provide sufficient samples of grammatical sentences which makes it

impossible to use them in the derivation of language structure. The following is his view on this matter:

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will so wildly skewed that the description would be no more than a mere list.
>
> Chomsky (1962, p. 159)

The way the language faculty is organised he added, cannot be based on mere statistical approximation or analysis of performance phenomena. His view and judgement of CL as the one quoted above influenced the course of linguistics enquiries in the late 1950s, putting CL, according to Leech (1992) "in the backwater where they were neglected ..." (p. 110).

In spite of these criticism and unpopularity, works on CL did not cease. Although, the volume was affected, they nonetheless had made tremendous contribution to the development and strength of CL. These works include Quirk's collection of written and spoken English in his Survey of English Usage (SEU) which began in 1959, followed by the development of Brown Corpus in 1963 led by the key figures Francis and Kucera. Brown Corpus consists of one-million-word sample of standard present-day American English and it was the first to be computerised. Quirk's SEU and the establishment of Brown Corpus marked the revival of CL. At the same time of the building of Brown Corpus, Sinclair at Edinburgh University established the first computerised corpus of spoken English (British). Soon after, large scale corpora such as London-Oslo-Bergen (LOB) corpus (1971), COBUILD Bank of English (1982) and British National Corpus (1995) were built. These large scale projects have made

remarkable contribution to the language research and development, but more importantly have transformed CL into a leading research paradigm.

Since then CL has expanded to an array of other linguistic areas; English for Specific Purposes (Conrad, 1996, 2001, 2002; Flowerdew, 1998, 2005) pragmatics (de Beaugrande, 1996), discourse (Baker, Gabrielatos & McEnery, 2013a, 2013b; Baker et al., 2008; Cheng & Lam, 2013; Marín Arrese, 2015; Pérez-Paredes, Jiménez & Hernández, 2017; Upton & Cohen, 2009; Upton & Connor, 2001; Wichmann, 2004), syntax (Yunus & Awab, 2011, 2012, 2014) and semantics (Ali, 2007) to name a few. The application of CL methodology enables linguists to define an area of linguistics and differentiate the approaches taken to language analysis, for example the distinction between corpus-based syntax versus non-corpus–based syntax or corpus-based discourse analysis versus non-corpus-based discourse analysis (McEnery & Wilson, 2001, p. 2).

In recent years CL has also been extended to learner language, giving rise to learner corpus research (henceforth LCR) (Gilguin & Granger, 2015). LCR, which grows in tandem with the establishment of various learner corpora worldwide, has contributed significantly to second language acquisition (SLA). Unlike the traditional SLA research, which predominantly emphasises on learner competence by characterising learners' underlying knowledge of the L2 (Ellis, 2008), LCR focuses exclusively on learner performance in the L2 by analysing and describing "the use of language by learners in actual production" (Gilguin & Granger, 2015, p. 1). The application of corpus linguistics tools and techniques has also enabled LCR to be conducted to a degree of automation, allowing analysis of whole learner population. Most importantly CL allows for investigation of authentic learner production as learner corpus compilation entails for language sample to be gathered from "genuine communication" (Sinclair, 1996) as a result of authentic classroom activities (Granger, 2002).

The understanding that CL is a powerful methodology enabling aspects of learner language which could be hidden to be revealed and described with corpus methodology, this study has adopted a corpus-based approach to studying learner language. The construct of this study is entirely data-driven with the use of computer learner corpus in obtaining the quantitative and qualitative data on the use of *BE* by the L1-Malay ESL learners. Corpus-based methodology requires for the selection of a set of linguistic categories prior to a study. For this study the researcher has identified the forms and functions of *BE* as the focus of the corpus investigation.

1.1 Statement of Problem

Many studies on second language acquisition have recorded variability in the production of *BE* among L2 learners of English. Ionin and Wexler (2001) Pine, Conti-Ramsden, Joseph, Lieven, & Serratrice (2008), Schütze (2004), Tode (2003) and Theakston & Rowland (2009) for instance reported that *BE* was frequently omitted in the early stages of child L2 acquisition. There is also evidence of ESL/EFL learners using *BE* in various ways not conforming to its standard forms and functions such as the use of *BE + bare V* or *BE + Ved* construction to mark tense and agreement as in *"The lion **is go** down"* (Ionin & Wexler, 2001). Researchers also reported that *BE* tends to be inserted before a special category of intransitive verbs; unaccusative verbs to produce *BE + Ven* construction as in *"What **was happened** yesterday"* (Ju, 2000; Oshita, 2000; Park & Lakshmanan, 2007; Yip, 1994). It is also common for *BE* to manifest itself as agreement errors as in "*Princess Isabella **are** very kind and gentle.*" (Maros et al., 2007) or to be replaced with another verb (e.g. auxiliary *do*/*does*) as in *"The children **do not** at home now*" (Unlu & Hatipoglu, 2012).

The same types of misuses of *BE* were also attested in the language production of ESL learners in Malaysia. *BE* was reported to be dropped before adjectival (*It **Ø** also good*

for- for our reading) (Ting et al., 2010), nominal (*My cat's* ***Ø*** *name Coco*) (Maros et al., 2007) and locative predicates (*..and my brother* ***Ø*** *also in Kedah*) (Platt & Weber, 1980), and overgenerated before a main verb to produce constructions such as "*The nurse* ***was bandaged*** *her leg*" (Wee, 2009) or "*My family and I* ***was go*** *to Pulau Tioman…*" (Arshad & Hawanum, 2010). In addition, learners also produced errors in tense "*I* ***am*** *so lucky because my family and I were save and nothing happen to us*" (Maros et al., 2007) and agreement "*These changes* ***is*** *depending on the current situation.*" (Siti Hamim & Mohd Mustafa, 2010).

The literature discussed, thus far, highlighted the difficulties ESL learners encounter in the correct use of *BE*, hence confirming that the verb is generally difficult for ESL learners. The tendency exhibited by Malaysian ESL learners to use *BE* incorrectly has ignited the researcher's interest to investigate the use of the verb more extensively. The decision to conduct an in-depth investigation on the use of *BE* is also motivated by the complexity of the forms and functions of the verb. The learners' competency in the use of *BE* or the lack of it could reveal many aspects of the target language mastery, which include competency in the morphological, grammatical and syntactical aspects of the target language. More importantly the analysis of *BE* would allow the researcher to link research to practice since the findings from the study would provide valuable input to improve second language teaching and learning practice generally and the teaching of *BE* to ESL learners specifically.

1.2 Motivation of the Study

A major motivation for this investigation is to foreground what learners can and cannot do with *BE*, rather than focusing solely on the errors that they make when using *BE* in their writing. As stressed by Ellis (2008) "we need to know what learners do correctly as well as what they do incorrectly" in order to "provide a complete picture of the

learner language" (p. 61). Empirical data on learners' grammatical use of *BE*, paired with the data of their ungrammatical use would enable researchers and educators to determine more accurately the degree of the learners' competency as well as their problems in the use of *BE*. Hence, they would be able to focus on the specific areas of difficulties and design more effective interventions that could cater to these specific areas of difficulties.

Another major motivation for this study is that despite the growing interest on *BE* worldwide, there has not been any similar attention given on the verb in the Malaysian research scene. Several error analysis studies conducted in Malaysia did include *BE* in their research analysis (Maros et al., 2007; Ting et al., 2010; Wee, 2009; Wee, Sim & Kamaruzam, 2010), but only as a part of other major errors in the L2 learners' data that they analysed or the analysis focused on a very specific *BE* morpheme (e.g. Arshad & Hawanum, 2010; Jishvithaa, Tabitha & Kalajahi, 2013; Manokaran, Ramalingam & Adriana, 2013). These studies did not give a full and comprehensive account of the use of the verb by the learners. Thus, there is a need for a research that can provide a comprehensive pattern of the use of *BE* in all its forms and major functions.

Thirdly, there are also varying degrees of focus given to *BE*, making it difficult to ascertain its overall behaviour in the L2 learner language repertoire. Some researchers placed *BE* as a part of a larger investigation into the acquisition of verbal inflection (Hawkins & Casillas, 2008; Haznedar, 2001, 2007; Haznedar & Schwartz, 1997; Prevost & White, 2000), morpheme studies tend to focus on either *BE* copula or *BE* auxiliary, treating each as a separate grammatical morpheme (Fleta, 2003; Moscati, 2006; Muneera & Wong, 2011; Theakston & Rowland, 2009; Tode, 2003, 2007), while error analysis studies have the inclination to treat *BE* only as a minor part of a larger investigation into learner errors (Maros et al., 2007; Murad & Khalil, 2015; Wee, 2009; Wee, Sim & Kamaruzaman, 2010). The lack of real focus on the verb has made it even

more challenging for researchers to draw a unified and holistic account of the factors influencing L2 learners' patterns of the use of the verb.

Maros et al. (2007), Wee (2009), Wee, Sim and Kamaruzam (2010), Lee and Huang (2004), Chan (2004), Muneera and Wong (2011) and Murad and Khalil (2015) named L1 transfer to be responsible for the ill-formed *BE* constructions. These researchers argued that the non-existent of copula-like verb in learners' native language is a major influencing factor. Rejecting the notion of negative interlingual transfer, Oshita (2000), Ionin and Wexler (2001, 2002), Fleta (2003), Herat (2005) and Unlu and Hatipoglu (2012) attributed the deviant constructions of *BE* to the developmental aspects of language learning and acquisition (Dulay & Burt, 1974; Dulay, Burt & Krashen, 1982). By conducting a thorough investigation on all aspects of *BE*, the researcher would be able to clearly identify the overall patterns of the use of all forms and functions of the verb the ESL data she analysed and this would help her to determine the exact factors influencing the grammatical use as well as ungrammatical use of *BE*.

Fifthly, a majority of the previous studies employed elicitation (e.g. Chan, 2004; Hawkins & Casillas, 2008; Ju, 2000; Lee & Huang, 2004) and longitudinal research frameworks (e.g. Fleta, 2003; Gavruseva & Meisterheim, 2003; Haznedar, 2001; Lakshmanan, 1995), which involved restricted number of participants. These studies have no doubt contributed significantly to the depth of the investigations, however, the findings could not be generalised across other L2 settings. The present study adopts a corpus-based methodology, which enables to a sizeable authentic data to be analysed. The advent of computer technology with the development of more sophisticated lexical analysis software (i.e. WordSmith Tools), enables various features of the learner language to be analysed concurrently with the speed and ease not possible with non-corpus-based methodology.

Finally, there exists a very wide gap in the literature concerning the learners' age and level of L2 exposure. Researchers have collected and analysed tremendous volume of child L2 learner data (e.g. Fleta, 2003; Gavruseva & Meisterheim, 2003; Haznedar, 2001; Hawkins & Casillas; 2008; Ionin & Wexler, 2001; Lakshmanan, 1995; Lee & Huang, 2004; Tode, 2003; Unlu & Hatipoglu, 2012), however less focus was given on more advanced L2 learners especially those who have received many years of formal instructions in the target language. The insights into the patterns of use and problems that this older and more advanced group of L2 learners is facing would be valuable to L2 research community in determining the natural course of L2 learners' acquisition of the target language and how far formal instructions have affected the development of learners' acquisition.

The existing problems arising from previous studies of *BE* and the lack of such investigation into the Malaysian ESL learner data call for a comprehensive and unified examination of *BE* in the ESL learners' language and this is a challenge that the present research intends to fulfill.

1.3 Purpose of the Study

This study aims to provide a comprehensive account on the use of *BE* by the L1-Malay ESL learners and from the findings propose a corpus consultation model for the teaching of *BE* in helping the learners improve their writing. The study focuses mainly on data from more advanced ESL learners, who have had more than 11 years of formal exposure to English.

1.4 Objectives of the Study

The research objectives of this study are as follows:

1. To compare and contrast the use of *BE* in the essays compiled in MACLE with the essays compiled in LOCNESS.

2. To analyse the forms and functions of *BE* in the essays compiled in MACLE and to provide a comprehensive account of the use of the verb in the learner essays.

3. To analyse if the syntactic environments influence the grammatical and ungrammatical uses of *BE*.

1.5 Research Questions

The study aims to answer the following research questions:

1. What are the similarities and differences in the use of *BE* in the essays compiled in MACLE and LOCNESS?

2. What are the distributional patterns for each form and function of *BE* in the essays written by L1-Malay learners in the Malaysian Corpus of Learner English?

3. What are the patterns of the (a) grammatical and (b) ungrammatical uses of *BE* in the essays written by L1-Malay learners?

4. How do the syntactic environments influence the grammatical and ungrammatical uses of *BE* in the essays written by L1-Malay ESL learners?

In addressing the "so what" question, this study also proposes a corpus consultation model for the teaching of *BE* to ESL learners in Malaysian universities as one of the efforts to improve the learners' writing. The findings from the analysis of *BE*, which would highlight the most challenging aspects in the use of the verb and the possible influence from the syntactic environments on the grammatical and ungrammatical uses

will be used as the foundation for the proposed model. The description and discussion of the corpus consultation model are presented in detail in Chapter 8 of this thesis.

1.6 Scope of the Study

MACLE consists of essays written by Malaysian ESL learners from three major ethnic groups; Malay, Chinese and Indian, therefore, the essays were written by learners from three different L1s Malay, Chinese and Tamil respectively. For the purpose of this research, only essays written by Malay learners will be analysed, mainly because the findings have to be discussed in relation to interlingual transfer. Previous studies have attributed the variability in the use of *BE* to negative interlingual transfer (Maros et al., 2007; Nor Hashimah et al., 2008; Wee, 2009; Wee, Sim & Kamaruzam, 2010). Malay, Chinese and Tamil are distinctively different structurally, thus, each may have different and varying effects on the learners' English. By excluding the essays written by L1-Chinese and L1-Tamil learners and narrowing the analysis to L1-Malay learners' essays enables for the positive and negative transfer of learners' L1 to be examined more closely. The researcher's proficiency in both spoken and written Malay would allow for the discussion of the extent of L1 transfer to be conducted with ease.

The study attempts to provide a comprehensive account of the use of *BE* in the learners' writing, thus, the analysis focuses primarily on all the *BE* forms; five (5) finite forms (*am, is, are, was, were*), three (3) non-finite forms (*be, been, being*) and the functions of the verb (copular, auxiliary, negative operator, interrogative operator, existential *there*, *it*-cleft). In order to identify the patterns of use and determine if the syntactic environments influence the use of *BE*, the constituents occurring before and after *BE* are also examined. These constituents include the types of subjects, subject predicates, form and class of post-*BE* lexical verbs and the presence of intensifiers and auxiliaries. *Longman Grammar of Spoken and Written English* (Biber et. al, 1999) is utilised as the

main reference for the setting up of the analytical parameters for the forms, functions and pre-*BE* and post-*BE* constituents for this study.

In order to provide a comprehensive picture of what learners can and cannot do with *BE*, the study analyses the correct and the incorrect uses of the verb. The incorrect use refers specifically to covert and overt *BE* that do not conform to the productive rules of the English grammar, thus, termed the ungrammatical use. In contrast, the correct use refers to *BE* that is used in accordance to the English grammar, thus, termed the grammatical use in this study.

1.7 Significance of the Study

This section discusses seven major significances of the study. First and foremost, the study adopts an entirely corpus-based approach, which allows not only for a larger learner data to be analysed, but also enables analysis of an extremely wider spectrum of *BE* use. In this study, the analysis encompasses all the grammatical and ungrammatical uses of *BE* in various forms and functions. This analysis provides a strong foundation for the mapping of *BE* forms to their functions and ultimately could unfold major and distinctive patterns of *BE* use among the L1-Malay ESL learners in Malaysia. In addition, since the corpus is a representation of authentic language the researcher will be able to investigate the learners' actual English language output, which will uncover not only their competency and development in the target language, but also the pertinent language problems they face. The findings from this study not only serve the purpose of designing curriculum and teaching materials for the major functions of *BE*, but will also provide SLA researchers with a comprehensive data of *BE* use among L1-Malay ESL learners in Malaysia.

Secondly, the study makes use of large data, something that would be impossible for a non-corpus-based study. The advent of computer technology in forms of concordancer

software and the establishment of large learner corpora worldwide provide an opportunity for researchers to investigate sizeable corpora of authentic learner data and at the same time administer comparative analysis of the different learner corpora. The corpus under investigation i.e. Malaysian Corpus of Learner English (MACLE) (Knowles & Zuraidah, 2004; Knowles et al., 2006) has approximately 800,000 word tokens. The analysis of such large data will generate a more reliable account of the state of ESL learners' language, thus making it possible for the findings to be generalised to other L2 settings. Ultimately the findings can also be a valuable reference for the formulation of learner language or interlanguage hypothesis.

Thirdly, this study will be the first to conduct an in-depth analysis of MACLE. Since its completion in 2004, the corpus has not been exhaustively investigated. To this date only five publications have been produced on MACLE (Aziz & Mohd Don, 2013; Aziz & Mohd Don, 2014; Knowles et al., 2004; Knowles & Zuraidah, 2006; Mohd Don & Sriniwass, 2017), two of which reported parts of the findings from the present study (Aziz & Mohd Don, 2013; Aziz & Mohd Don, 2014). This study is the first to embark on a comprehensive investigation on the corpus data and it could become a valuable source of reference for future research on MACLE.

Fourthly, this study is also going to be the first to provide a corpus-based account of the L1-Malay learners' use of *BE*. The findings can be compared to existing findings from studies employing different methodological frameworks such as error analysis or contrastive analysis. Moreover, data from this study are also comparable to other NNS or NS learner corpora. Such comparison will present a fuller understanding and deeper insights into ESL learners' language behaviour or the systems inherent in it.

Fifthly, this will be the first in-depth study conducted on the patterns of L1-Malay ESL learners' use of *BE*. Error analysis investigations conducted for instance by Maros et al.

(2007), Wee (2009), Wee, Sim and Kamaruzam (2010) and Ting, Mahanita and Chang, (2010) did include some discussions on the variability in the supply of *BE*, however, since these studies were interested in identifying and describing major grammatical errors (which *BE* was a part of) they did not provide an exhaustive account of either the ill-formed or well-formed *BE*. This study proposes to fill this research gap by providing a detailed account of the use of both grammatical and ungrammatical uses of all forms and functions of *BE* and provide a comprehensive patterns of the overall behaviour of *BE* in the language data of more advanced ESL learners in Malaysia.

Sixthly, and perhaps the most significant contribution of this study is that it not only focuses on learner errors, but it also analyses the grammatical use of *BE* in the learner corpus. At present, there has not been any study in Malaysia that focuses on both the grammatical and ungrammatical uses of *BE*. Hence, the findings from this study would be the first to supply such data. The findings would unveil the patterns of use for all the forms and functions of *BE*. These patterns can reveal many aspects of language acquisition including issues on L2 acquisition process and development, factors affecting acquisition, learnability and ESL language teaching and learning.

Finally, the study is also significant in proposing a corpus consultation model for the teaching of *BE* to ESL learners in Malaysian universities. Based on the findings of the patterns of the grammatical and ungrammatical uses of *BE*, the study shall propose a corpus-based teaching and learning model for *BE*. A substantial number of corpus studies conducted have reported successful implementation of corpora in the language classroom and most importantly they have also provided empirical evidence of the effectiveness of this approach in teaching various aspects of language including lexico-grammatical patterns, grammar, vocabulary, collocations (Chambers & O'Sullivan, 2004; Leel, 2011; Miceli & Kennedy, 2002; Phoocharoensil, 2012; Shi, 2017; Todd, 2001; Vannestal & Lindquist, 2007; Yunus & Awab, 2012, 2014) and writing

(Chambers & O'Sullivan, 2004; Gaskell & Cobb, 2004; Kennedy & Miceli, 2001; Kotamjani, Razavi & Hussin, 2017; Miceli & Kennedy, 2002; O'Sullivan & Chambers, 2006; Yoon, 2008; Yoon & Hirvela, 2004). In general, corpus-based instructional technology has great potentials in second language teaching and learning and in view of these potentials this study shall propose an integration of corpus in the teaching and learning of *BE* to ESL learners in Malaysia. It is hoped that this study would provide language teachers with some ideas and guidelines to implement the approach in teaching of not only *BE*, but also other linguistic aspects.

1.8 Structure of the Dissertation

This dissertation comprises of eight (8) chapters. This chapter consists of an overview of the study, which includes the explanation on the choice of *BE* as the focus of the study, the purpose, objectives, research questions, scope and the significance of the study.

Chapter 2 presents the literature review, which focuses on the theoretical, methodological and empirical backgrounds of the study. It first provides an overview of corpus linguistics and computer learner corpora. The section also reviews and discusses the major approaches to learner corpora research. Based on these discussions, the research framework adopted for this study is outlined. Finally, the chapter reviews past studies on *BE* and summarises the inadequacies of research on the use of *BE* by L2 learners in Malaysia.

Chapter 3 describes in detail the methodology adopted in answering the research questions. It first introduces the corpora utilised for the study, followed by detailed and comprehensive description on the coding, retrieval and analysis processes.

Chapter 4 presents the findings from the quantitative analysis, which encompasses mainly the descriptive statistics on the forms and functions of *BE*, the patterns of the

grammatical and ungrammatical uses and the influence of syntactic environments on the use of *BE*.

Chapter 5 presents the qualitative findings, providing the textual evidence of how learners use *BE*. The findings are presented in relation to the patterns of use, influence of syntactic environments and syntactic complexity.

Chapter 6 summarises and discusses the major findings of the study. The phenomena discovered with regard to the grammatical and ungrammatical uses of *BE* are also explained in the chapter.

Chapter 7 concludes the thesis by presenting the major findings of the study and discussing the implications of the study and its limitations. It ends with several suggestions for future research.

Chapter 8 outlines and proposes a corpus consultation model for the teaching of *BE* to the ESL learners in Malaysian universities. It begins with an overview of corpus consultation in the second language teaching and learning, followed by a detailed description of the corpus consultation model.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

Chapter 2 reviews the historical issues, theoretical issues and methodological issues related to the study. It includes examining past work that has been done in this area, past and present frameworks of explanation that have been used and how other researchers dealt with similar (or even different) problems. Relevant literature on the incorporation of corpora in language teaching will also be reviewed. The inclusion of the pedagogical aspect of corpora is motivated by the need to propose a model for the incorporation of corpora in the teaching of *BE* to ESL learners in Malaysia. This will provide the answer to the "so what" question.

2.1 Methodological Background

This study adopts the corpus linguistic (CL) methodology, which characterises itself on a set procedures or methods for studying language (McEnery & Hardie, 2012, p. 1). According to Tognini-Bonelli (2001) it is an empirical approach, which makes use of authentic data and aims to describe language as it is realised in text(s) (p. 2). This section focuses on reviewing the major approaches to analysing language within CL methodology as well as past studies, which have utilised these approaches.

2.1.1 Approaches to Corpus Analysis

The CL methodology distinguishes two distinct approaches to language analyses, corpus-based approach and corpus-driven approach. The terms were originally introduced by Tognini-Bonelli (2001) to differentiate the approaches undertaken to analyse corpus data. This section discusses the approaches further.

2.1.1.1 Corpus-based Approach

Corpus-based approach is referred to "a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that are formulated before large corpora became available to inform language study" (Tognini-Bonelli, 2001, p. 64). It lends support to the linguistic theories deriving from the reflection of linguistic scholars, as such a corpus would provide the quantitative evidence to validate the existing theories or at times even indicate where minor or major adjustment and correction to be done to any theory or model adopted in a particular linguistic enquiry. In a corpus-based enquiry the corpus evidence is merely used as the supporting data to validate a pre-existing set of categories. Biber (2009) further explained that a researcher adopting this approach would generally begin by selecting a set of linguistic categories as a priori to the investigation and utilising corpus methodology to "describe the patterns of variation and use associated with those grammatical feature" (Biber, 2009, p. 278). The present study emulates this principle. It begins by identifying the linguistic category it intends to analyse that is *BE*.

Corpus-based studies have provided new insights into many areas of language structure and use (Biber, Conrad, & Reppen, 1994). Corpus-based analysis is an empirical approach, which includes the study of language as it is used in natural context and the analysis involves the use of specialised computer software. An important feature of corpus-based analysis is that it recognises the interdependence of lexis and grammar and the balance between routine and creativity in language use (Stubbs, 1993). Biber et al. (1988) summarised corpus-based analysis into four essential characteristics:

i. It is empirical, analysing the actual pattern of use in natural texts.
ii. It utilises a large and principled collection of natural texts, known as "corpus" as the basis for analysis.

iii. It makes extensive use of computers for analysis, using both automatic and interactive techniques.

iv. It depends on both quantitative and qualitative analytical techniques.

Biber, Conrad and Reppen (1998, p. 4)

By employing a corpus-based analysis, researchers are able to provide accurate descriptions of naturally occurring language in terms of the frequency of occurrence of particular content and function words, clusters, formulae, lexical bundles or linguistic features. The findings from these investigations provide insights into the authentic use of language as opposed to the intuition of what native speakers have on what should occur in natural language. The insights can be used to reassess the order of the teaching of the grammatical system and also to reevaluate the focus of certain grammatical aspects in language syllabus and textbooks. Biber et al. (1998) highlighted three advantages of the corpus-based approach:

i. by including a relatively large collection of texts, corpus-based analyses provide the basis for generalisations concerning groups of speakers and writers at different stages,

ii. by investigating the association patterns among sets of linguistic features, corpus-based analyses enable more comprehensive descriptions of language use at different developmental stages, and

iii. by including texts from multiple registers, corpus-based analyses facilitate broader perspectives on language development.

Corpus-based studies have been acclaimed for their success in describing and providing comprehensive accounts of language variation, such as Biber's (1986a, 1986b, 1988) seminal investigations into the linguistic variation between spoken and written register of English. Employing multi-dimensional comparison, Biber highlighted prominent

linguistic characteristics of different language registers; spoken and written, from a wide range of texts; medical research articles (Biber & Finegan, 1988) and university prose (Biber, Conrad, Reppen, Byrd & Helt, 2002). In addition, work done on CANCODE project, particularly those conducted by Carter and McCarthy (1995), Hughes and McCarthy (1998) and McCarthy (1998) are also excellent examples of corpus-based studies on linguistic variation. These studies have successfully unraveled the existence of specific linguistic features in the spoken variety (pre-posed and post-posed items and several kinds of ellipsis), which were not represented in the grammar of written register (Carter & McCarthy, 1995; Hughes & McCarthy, 1998; McCarthy, 1998). Corpus-based investigations on language variation were not only dominated by examination on spoken and written language varieties but they also included studies on the variation between and within academic prose such as the work done by Conrad (1996) on the differences and similarities in the language used in two types of biology texts; ecology textbooks and professional research articles in ecology.

Besides successful application in language variation studies, corpus-based analysis has equally made its mark in lexis studies. Many analyses of individual words or phrases making use of concordance output were conducted. Typically these types of analyses involved careful examination of the collocation patterns, which allowed a researcher to draw on what Louw (1993) referred to as the 'semantic prosody' of the lexis and their connotation. Many such works were directly associated with the compilation of modern English dictionaries. Channell (2000) for instance, while working on the compilation of Collins COBUILD English Dictionary, analysed the evaluative function of words and expressions by systematically examining the collocation patterns existing from concordance lines. Her findings revealed that pragmatic meaning was often hidden from introspection, arguing that the analysis of evaluation can be done more

systematically, thus, more accurately using the corpus-based approach she employed, rather than relying on chance and intuition.

Another notable work on lexicography adopting a corpus-based approach, was the work of Biber, Johannson, Conrad, Finegan and Leech (1999) in producing *Longman Grammar of Spoken and Written English*. It is a corpus-based grammar coursebook and reference, which has made use of the Longman Spoken and Written English Corpus (LSWEC) as its reference corpus. Unlike other traditional grammar texts, it is the first grammar text to be based on large and balanced authentic spoken and written texts. LSWEC has over 40 million word tokens, which provides a very sound basis for analysis of both written and spoken grammatical patterns.

Other than lexis, there is also countless number of studies investigating specific grammatical features that have adopted the corpus-based approach. Biber (2003) for instance, analysed compressed noun-phrase structures in newspaper discourse from LSWEC (Biber et al., 1999). By counting the frequency of occurrences of the selected noun-phrases and comparing the frequency to those occurring in other registers, namely conversation, fiction and academic texts, Biber (2003) managed to obtain unique and interesting usage of compressed noun-phrase.

The recent development in CL has also seen corpus-based approach being extended to textlinguistics; specifically it has been applied to investigate for instance functional and rhetorical aspects in specialised corpora. Upton and Connor (2001), for example integrated corpus-based methodology with traditional genre analysis in their cross-cultural study of politeness strategies in job application letters. Wichmann (2004) employed corpus-based methodology to investigate the prosodic aspects of *please*-request. Using the data from the spoken corpus in International Corpus of English (British contribution), the researcher examined in detail the intonation contours of

please with respect to its position, speech acts and context. A more recent study employing corpus-based methodology on textlinguistics includes a study conducted by Torgersen, Gabrielatos and Hoffmann (2011) on pragmatic markers in London English. The investigation on two corpora of spoken London English (Linguistic Innovators Corpus and Corpus of London Teenage Language) revealed variation in the use of pragmatic markers according to sex, ethnicity and geographical location. Most recently, Sabet and Minaei (2017) conducted a comparative corpus-based analysis on different parts of quantitative and qualitative research articles in the field of TEFL. Applying corpus methodology in the conventional discourse analysis framework, the study successfully revealed differences between quantitative and qualitative research articles from lexico-grammatical and rhetorical features. Other corpus-based studies on discourse are reviewed in Section 2.1.2.1 and they include Aziz, Jin and Nordin (2016) on interactional metadiscourse and Mohd Don and Sriniwass (2017) on conjunctive adjuncts.

As exhibited by the studies reviewed, corpus-based methodology has expanded beyond mere frequency and collocation investigations. It has made remarkable contributions to other areas of linguistics, when studies in pragmatics, discourse or semantics incorporate this methodology alongside the conventional analysis specified for each linguistic discipline. The application of corpus-based methodology in tandem with the advent of computer technology has made it possible for researchers to analyse larger data ensuring, therefore, the data are representative of the discourse community or language users under investigation. In addition, this collaboration enables multi-level, thus, in-depth analysis to be administered that could yield a more comprehensive description of a language phenomenon and ultimately provide an empirically sound foundation for the derivation of language hypotheses.

2.1.1.2. Corpus-driven Approach

Corpus-driven approach primarily aims to draw linguistic categories based on systematic analyses of recurrent patterns and distributional frequency inherent in the corpora investigated (Tognini-Bonelli, 2001). In essence, the approach takes an exploratory or inductive move towards language studies. Instead of beginning with a priori of linguistic categories, it sets out with the exploration of the language in context, systematically deriving the evidence for linguistic categories through the emerging recurring patterns and frequencies existing in the corpora. Interestingly as pointed by Tognini-Bonelli (2001), the approach opens up grounds for the discovery of new language hypotheses or theories that may not necessarily support the existing ones. Sinclair (1991) posited the interdependence of evidence to the theoretical statement made by studies done employing a corpus-driven approach. This is to highlight that the derivation of theories is based on corpus evidence, where examples are taken verbatim and never ignored or adjusted.

A significant and major utilisation of the corpus-driven approach was evident in the work done by Hunston and Francis (2000) on the 'pattern of grammar' framework that focused principally on grammatical patterns through inductive analysis of the corpus (Biber, 2009). Through the study Hunston and Francis (2000) managed to successfully unveil the "systematic regularities in associations between grammatical frames, sets of words, and particular meanings on such a larger scale than it could have been possible to anticipate before the introduction of large-scale corpus analysis" (in Biber, 2009, p. 278). Biber (2009) pointed out that most lexical collocations studies, for instance by Biber (2009), McEnery et al. (2006) and Partington (1998) can be regarded as corpus-driven as the collocation patterns were derived entirely from corpus analysis (Biber, 2009).

In a stricter sense of its application, a corpus-driven methodology in the lexical collocation studies generally or formulaic language specifically would have to possess the following characteristics:

i. It would be based on analysis of the actual word forms that occur in the corpus (not lemmas).

ii. It would be based on analysis of sequences of word forms, with no consideration given to the grammatical/syntactic status of those words.

iii. It would focus on frequent, recurrent combinations of words.

Biber (2009, p. 281)

Having reviewed the radical aspect of corpus-driven methodology, Biber (2009) cautioned against the misconception of its superiority in the light of corpus-based approach, as he eloquently pointed out, each involves "radically different methods" and can both be potentially powerful tools to unravel "different perspectives of language structure and use" (p. 279).

More recently corpus-driven approach has moved beyond lexical collocations and the approach has been applied quite extensively in multimodal and critical discourse analysis. Works conducted by Baker, Gabrielatos and McEnery (2013a, 2013b) on the representation of Islam and Muslim in the British press are excellent examples of the collaboration of corpus-driven approach with critical discourse analysis. The methodology undertaken enables the researchers to reveal more extensively how Muslim and Islam are represented in the western media in particular in the British press. The studies reported that Muslims were most frequently associated with conflicts and represented as easily offended, alienated and in conflict with non-Muslim.

Pérez-Paredes, Jiménez, and Hernández (2017) following the footsteps of Baker et al. (2013a, 2013b) examined the representation of immigrants the British legislation and administration informative texts from 2007 to 2011. By complementing collocational analysis with CDA methods, the research was able to shed some light on how the British immigration laws and official texts construct the image of immigrant population the country. The study found that even though the British administration avoided explicit negative construction of immigrants coming to the country, the immigrants were "partially constructed as homogeneous, well-categorised group through very limited set of lexical item" (p. 81). The researchers argued that the representation during the period investigated were more focused on legitimasing the control over the immigrants rather than on establishing better immigration policies to accommodate these groups of individuals.

Other prominent works adopting the integration of corpus-driven method with CDA methods include Cheng and Lam (2013) who investigated Western perceptions of and relations with Hong Kong after the handover from British to China in 1997; Chan, Albakry, Williams, Lamb, Kelsey, van Dijk and Owens (2017) who examined the representation of Umbrella Movement in newspaper in Hong Kong and in China; Brindle (2015) who studied the representation of the Sunflower student movement in the Taiwanese press; and Marín Arrese (2015) who analysed the epistemicity and stance in English and Spanish journalistic discourse. These studies have contributed significantly to linguistic research in general, as they have managed to identify the presence and distributions of a wide range of patterns of language use from large quantities of news articles and the journalists' preferences for certain grammatical patterns and lexical choices when referring to particular issues, individuals and events (Chan et al., 2017, p. 5). More importantly, they have widened and diversified the

scope of corpus-driven approach to suit the needs and atmosphere of linguistics in the recent times.

2.1.2 Computer Learner Corpora

In the tradition of corpus linguistics, investigation on learner corpora came into being quite later; the interest started to emerge in the late 1980s and early 1990s when researchers began efforts to compile written data of non-native learners of English. Computer learner corpora (CLC) to borrow to Sinclair's (1996) definition can best be described as:

> ... electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.
>
> Sinclair (1996)

As postulated by Sinclair (1996), corpora have to be collections of real communication taking place in the natural surrounding where people go about their everyday lives. In the context of NNS learner corpora, the term 'authentic'; due to the nature of where and when English is used by the ESL/EFL learners refers to written essays which may be in the form of free or elicited essays. Granger highlights the fact that in the case of learner corpora, the compilation is "rarely fully natural", however, it is the result of "authentic classroom activities" (Granger, 2002, p. 8).

To qualify as a learner corpus, the data must originate from the non-native variety of English, which Granger (2002) has classified into English as an Official Language (EOL), English as a Second Language (ESL) and English as a Foreign Language (EFL), however, only the last two are placed under the umbrella of learner corpora. Another

important criterion of learner corpora is the data must encompass "continuous stretches of discourse" and "not isolated sentences or words" (Granger, 2002, p. 8).

Sinclair's (1996) definition of learner corpora specifies the need for the corpora to be built upon explicit design criteria, which normally details out the learner and task variables as the followings:

Table 2.1: Learner Corpus Design Criteria

LEARNER	TASK SETTING
• Learning context • Mother tongue • Other foreign language • Level of proficiency • […]	• Time limit • Use of reference tools • Exams • Audience/interlocutor • […]

Sinclair (1996)

In ensuring comparability with other learner corpora or native speaker corpora, it is important for the data in a learner corpus to be annotated with standardised annotation software. The learner and task variables for each text need to be documented and made available in the form of SGML file or separate files that are linked to the text file. Such documentation will enable researchers to easily expand the existing corpus by adding sub-corpus or conduct comparative analysis between the variables within the same corpus or with other corpora or sub-corpora.

Perhaps the most accessible and among the first compiled learner corpora, is the International Corpus on Learner English (ICLE) (Granger, 1998a), a collection of over two million words of EFL writings of learners from various L1 backgrounds; Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish and Swedish (Granger, et al., 2002). Taking the lead from Granger (1998a) other researchers and academicians from other parts of the world began their effort to compile their own learner corpora, comprising normally the data available in their own institution/country

setting. Some available learner corpora are Hong Kong University of Science and Technology Learner Corpus (HKUST), the biggest non-commercial corpus, thus far, with over 25 million words, TeleNex Student Corpus originating also from Hong Kong with 3 million words, Chinese Learner English Corpus (CLEC) currently with 1.2 million words and Uppsala Student English Project (USE) with 1 million words (Nesselhauf, 2004).

In Malaysia, several learner corpora have been established, following the need for such corpora to be used primarily in the research context of English as a Second Language. Among the currently available corpora are English of Malaysian School Students (EMAS) (Arshad, 2002) with 500,000 words, Malaysian Corpus of Learner English (MACLE) (Knowles & Zuraidah, 2004; Knowles et al., 2006) with 800,000 words and Corpus Archive of Learner English in Sabah-Sarawak (CALES) (Botley, De Alwis, Metom, & Izza, 2005) with 400,000 words.

2.1.2.1 Learner Corpus Research

According to Gilquin and Granger (2015) learner corpus research (henceforth LCR) is a new strand of research which focuses on language performance rather than competence. Its main objective is to describe the actual use of language by learners, which is made possible with the application of corpus linguistic tools and techniques (Gilquin & Granger, 2015). LCR has since been "characterised by the strong focus on lexis, lexico-grammatical and discourse phenomena" (Gilquin & Granger, 2015, p. 2). Some of the studies on lexis and lexico-grammatical aspects include Granger and Rayson (1998) on lexical profiling, Altenberg and Granger (2001) on lexical patterning of *make*, Nesselhauf (2003) on collocations, and Granger, Paquot and Rayson (2006) on multiword units. LCR studies on discourse include Hyland and Tse (2004) on metadiscourse, Gilquin and Granger (2015) on discourse markers, Hasslegard (2009) on

stance markers and Adel (2008) on involvement features. This section shall briefly review some of these representative studies as well as some recent studies conducted on learner corpora from Malaysia.

Granger and Rayson (1998) investigated the possibility of identifying stylistic characteristics of learner interlanguage using fully automatic lexical profiling software on POS-tagged data. The study made use of lexical frequency software developed at Lancaster University (Rayson & Wilson, 1996) on two similar-sized POS-tagged native and non-native learner corpora. LOCNESS was chosen as the control corpus, while the non-native learner corpus was drawn from the French component of ICLE. In order to reveal the distinctive features of French learners' interlanguage, the researchers compared nine major word categories (e.g. pronouns) and fourteen subcategories (e.g. personal or indefinite pronouns) in the learner corpora. The comparison had generated major features in the French learner writings, which according to researchers were typical of speech than of academic writing. The study demonstrates the value of POS-tagged data in LCR as well as the value of fully corpus-driven analysis of learner data. More importantly it is also the first to use fully automated software in identifying the stylistic features of learner interlanguage.

The study on collocation by Nessalhauf (2003) focused on the use of verb-noun collocation such as *fail an exam* or *take a break* among advanced learners of English in Germany. Manually extracted raw texts from the German component of ICLE were used. Collocations were analysed according to their degree of restriction and acceptability. The researcher found that about 25% of the collocations contained one or several mistakes caused mainly by wrong choice of verb and the problem was traced back to negative interlingual transfer. The findings suggest that collocations still present a major problem to advanced learners and that learners' L1 has a very significant effect on the incorrect use of collocations.

At the discourse level, the focus of LCR is mainly on the use of discoursal devices (e.g. discourse markers, stance markers or metadiscourse) following the use, underuse and overuse approach to data analysis (Barlow, 2005; Granger, 2002; Leech, 1998). Nevertheless, there are also studies which did not prescribe to this approach, one of which was by Hyland and Tse (2004) on the use of metadiscourse in L2 postgraduate writing. The aim of the study was to discover how student writers perceived and engaged with their disciplines through the use of interpersonal features of texts (metadiscourse). The corpus used in this study consisted of 240 masters and doctoral dissertations written by Hong Kong Chinese students. The study employed both qualitative and quantitative approaches to data analysis. In order to gain insights into the text data, semi-structured interviews with 24 postgraduates were conducted. The researcher found writers used slightly more interactive than interactional metadiscourse, and there was variation across the two degree corpora (Masters and Phd) as well as across disciplines. The study raises an important methodological statement that LCR does not always have to restrict itself to a one methodological framework (use, underuse and overuse). It also demonstrates the importance of complementing the corpus data with other data (e.g. interview), as they would provide the explanation to the descriptive data obtained from the corpus-based analysis.

Among the more recent LCR on discourse was by Gilquin and Granger (2015) on the use of two-word discourse markers (DMs) among non-native learners representing twelve (12) L1 groups. The non-native learner data were obtained from LINDSEI (Gilquin et al., 2010), while LOCNEC (de Cock, 2004) was selected as the control corpus. The main aim of the study was to find out how learners from different L1 backgrounds use DMs and compare that to the native speaker data. It also aimed at demonstrating the importance of analysing individual data in addition to pooled data. Findings from quantitative analysis revealed a general underuse of DMs among learners

and the use varied depending on learners' L1. Qualitative findings revealed aspects of DMs use that were hidden from the quantitative analysis. French learners underused *you know* but the use resembled the norm of the native speakers, compared to Polish-speaking learners, who overused the same DM but in the behaviour less similar to that of the native speakers. This study highlights the importance of complementing quantitative analysis with a qualitative one, as it could reveal aspects of language use that are hidden from the aggregate data; the term used by Gilquin and Granger (2015) to refer to data/findings obtained from quantitative analysis.

Gilquin and Granger (2015) argued that aspects of grammar had been under-researched in LCR mainly because investigations on grammatical features would ideally require the data to be POS-tagged or parsed. Thus, most of the LCR on grammatical features had the tendency to involve linguistic item that can be studied using raw data such as on causative *make* (Gilquin, 2012), *what*-clefts (Callies, 2009), modal auxiliaries (Aijmer, 2002) and demonstrative pronouns (Petch-Tyson, 2000). Nevertheless, there have been attempts by the LCR community to investigate grammatical features using annotated data and these studies commonly focus specifically on learner errors.

In 2003, Granger and her team at Louvain collected and error-tagged French Interlanguage Database (FRIDA) as part of FreeText project, which aimed at producing a learner corpus-informed CALL program for French as a Foreign Language. The project was based on the computer-aided error analysis developed for English by Dagneaux, Denness and Granger (1998). It adopted and combined Dulay, Burt and Krashen's (1982) descriptive error taxonomies (linguistic category and surface structure errors) and added another layer of information on errors to produce a three-dimensional error taxonomy. The team managed to error-tag a large proportion of the corpus (300,000 words) and manually detected and corrected 46,241 errors. The error-tagged learner corpus had contributed greatly to the development of FreeText CALL

programme and provided researchers with immediate access to detailed error statistics, which were of unparalleled importance in revealing areas of difficulty. Perhaps, the most important contribution of the study is that it paves the way to systemise error annotation, an area until today is still regarded as fuzzy due to the subjectivity of error taxonomy (Tono, 2003).

More than a decade after Granger (2003), LCR on learner interlanguage errors has still not expanded to its fullest potential, mainly because there has been very little success in devising a fully automated error-tagging system. In the study on L2 accuracy developmental patterns by Thewissen (2013), learner errors were still annotated manually by employing the Louvain error-tagging system (Granger, 2003). Computer-aided error analysis, which makes use of error-tagged corpora, is less popular in LCR as every stage of the method entails a number of problems (Gilquin & Granger, 2015). The identification stage is time-consuming since it is commonly conducted manually; it also involves deciding between real errors and infelicitous forms. There is also the issue of validity and reliability when tagging is done manually (Tono, 2003). In ensuring consistency in the assigning of tags, the tagging system has to be thoroughly documented. Tagged-errors also require correction and this stage is also problematic as it is not easy to reconstruct what the learner meant to say and there may be several plausible corrections to an error (Gilquin & Granger, 2015).

2.1.2.2 Learner Corpus Research in Malaysia

LCR in Malaysia has started to gain momentum in recent years, with the scope of the investigations gradually expanding. It grows in tandem with the establishment of several major learner corpora in the country, namely the English of Malaysian School Students corpus-EMAS (Arshad, 2002) and Corpus Archive of Learner English in Sabah/Sarawak-CALES (Botley et al., 2005), Malaysian Corpus of Learner English-

MACLE (Knowles & Zuraidah, 2004; Knowles et al., 2006) and Malaysian Corpus of Students' Argumentative Writing-MCSAW (Mukundan & Kalajahi, 2013). Earlier works on LCR involved mostly EMAS corpus (Arshad, 2002), since it was the first learner corpus to be established in Malaysia. EMAS is a collection of 800 learner compositions from three age groups (Primary 5- 11 years old, Secondary 1- 13 years old, and Secondary 4- 16 years old). One of the earlier studies includes Arshad's (2004) investigation on school students' language development in terms of vocabulary use and language productivity. More recent studies utilising EMAS include Arshad and Hawanum (2010) on auxiliary *BE*, Hong, Rahim, Hua and Salehuddin (2011) on verb-noun collocation and Kamarudin (2013) and Zarifi and Mukundan (2014) on phrasal verbs.

Another learner corpus facilitating LCR in Malaysia is CALES, which comprises of essays written by Diploma and Degree students from public universities in the states of Sabah and Sarawak. CALES was explored for spelling errors (Botley & Dillah, 2007), the use of idioms (Botley, 2010), evidence of L1 transfer in ESL learner writings (Botley, Haykal & Monalisa, 2005), and most recently on argumentative structure in ESL learner essays (Botley, 2014). The most recent learner corpus developed for investigating ESL learner language in Malaysia, MCSAW has also been actively contributing to LCR. MCSAW consists of argumentative essays written by secondary and college students in Malaysia. Studies utilising the corpus include Jishvithaa, Tabitha and Kalajahi (2013) and Manokaran, Ramalingam and Adriana (2013) on auxiliary *BE*, Arjan, Abdullah and Roslim (2013) and Loke, Ali & Zulkifli Anthony (2013) on prepositions, and Abdul Kader, Begi and Vasegi (2013) and Mukundan, Saadullah, Ismail and Jusoh Zasenawi (2013) on modal verbs.

Most LCR in Malaysia concentrates mainly on examination of specific grammatical feature in the Malaysian learner language. They include studies on auxiliary *BE* (Arshad

& Hawanum, 2010; Jishvithaa, Tabitha & Kalajahi, 2013; Manokaran, Ramalingam & Adriana, 2013), phrasal verbs (Kamarudin, 2013; Zarifi & Mukundan, 2014), prepositions (Arjan, Abdullah & Roslim, 2013; Loke, Ali & Zulkifli Anthony, 2013), modals (Abdul Kader, Begi & Vasegi, 2013; Mukundan, Saadullah, Ismail & Jusoh Zasenawi, 2013), lexical verbs (Kanestion, Singh, Shamsudin, Isam, Kaur & Singh, 2016) and articles (Abdul Rahim, Abdul Rahim & Chia, 2013).

The studies on auxiliary *BE* conducted by Arshad and Hawanum (2010), Jishvithaa, Tabitha and Kalajahi (2013) and Manokaran, Ramalingam and Adriana (2013) adopted the computer-aided error analysis framework (Granger, 2002) with their main focus on identifying and classifying learner errors in the use of auxiliary *BE*. Arshad and Hawanum (2010) analysed primary school student compositions extracted from the EMAS corpus, while the other two studies analysed argumentative essays in MCSAW. Manokaran et al. (2013) and Jishvithaa et al. (2013) focused mainly on identifying and classifying errors in the use of auxiliary *BE* in the past tense (Manokaran et al., 2013) and present tense (Jishvithaa et al., 2013), while Arshad and Hawanum (2010) analysed all types of errors committed with auxiliary *BE* since the focus of the research was not only to identify the sources of the errors, but also to propose possible teaching solutions to overcome specific problems students encounter with auxiliary *BE*. These computer error analyses have managed to identify and classify errors in the use of auxiliary *BE* by the Malaysian ESL learners among them include tense shift, agreement, missing auxiliary *BE*, wrong verb form, addition and misformation and misordering.

Other than *BE*, learner essays were analysed for the use of lexical verbs (Kanestion et al., 2016). Kanestion et al. (2016) conducted a preliminary corpus-based analysis of the use lexical verbs in Band 5 and Band 3 argumentative essays written by pre-university students from a pre-university college in Malaysia. The corpus consisted of twelve argumentative essays. The essays were POS tagged using the CLAWS 7 tagsets and a

concordancer was utilised to analyse the tagged data. The findings revealed four commonly used types of lexical verbs, namely past tense (VVD), *-ing* form (VVG), past participle (VVN) and *–s* form (VVZ). There was also no significant difference in the types of lexical verbs use in both Band 5 and Band 3 essays.

Interest has also been invested on the use of phrasal verbs (Kamarudin, 2013; Zarifi & Mukundan, 2014). Kamarudin (2013) in her investigation of the use of six phrasal verbs with particle *UP* in the EMAS corpus compared the Malaysian learners' use of phrasal verbs to that of the native speakers' from Bank of English (BoE) corpus. The learner data were POS-tagged using CLAWS tagger so that the phrasal verbs can be easily extracted using WordSmith Tools. The findings revealed that wrong usage of common phrasal verbs (e.g. *pick up, wake up, get up*) has strong association with the learners' lexical knowledge, their awareness of common collocates, familiarity with the context of use and most important their mother tongue. The appropriateness in the use of phrasal verbs was also found to improve over time, suggesting that learners had benefited from longer exposure to the target language.

Zarifi and Mukundan (2014) conducted a corpus-based analysis of the creativity and unnaturalness in the use of phrasal verbs among Malaysian ESL learners. WordSmith Tools 4.0 was used to extract the phrasal verbs, which then were tagged and lemmatised. The acceptability of the phrasal verbs used or created by learners was judged with the help of dictionaries and those without dictionary entry were judged against BNC. Learners were found to use phrasal verbs quite sparingly, but some of the phrasal verbs created were unnatural. Unlike Kamarudin (2013), this study did not compare the learner data with a reference corpus.

Arjan, Abdullah and Roslim (2013) investigated the use of prepositions (*in, on*) in argumentative essays written by secondary school and college students in Malaysia

drawn from MCSAW. Descriptive statistics were applied to determine the distribution and common errors of *in* and *on* as prepositions of place. Learners were found to be confused with the use of both *in* and *on* and had the tendency to mistaken one for the other. Nevertheless, there was a marked improvement in the accuracy of use as the learners progressed in their level of studies; college students recorded better performance than secondary school students.

MCSAW was also utilised for investigating the prepositions of time *on* and *at* by Loke, Ali and Zulkifli Anthony (2013). The investigation focused on determining the distribution of use and the common errors in the use of the prepositions. The study reported very limited use of both *on* and *at* as prepositions of time in the student essays, which the researchers explained was the result of the genre of the essays that limits the use of prepositions of time. The common errors recorded included omission, addition and incorrect use of preposition. Like other studies involving MCSAW this study did not detail out the methodological aspect of data annotation and analysis, which is crucial in corpus-based study on grammatical aspects. It also did not provide extensive samples of how learners use the prepositions in context, which could shed more understanding on how and why learners use these preposition and in what context they were correctly or incorrectly used.

Another grammatical item investigated using MCSAW was modals (Abdul Kader, Begi & Vasegi, 2013; Mukundan, Saadullah, Ismail & Jusoh Zasenawi, 2013). Abdul Kader, Begi and Vasegi (2013) compared the use of modals in argumentative essays written by secondary and college students in Malaysia, while Mukundan et al. (2013) examined the use of modals at the syntactic level in college student essays and determined the grammatical accuracy and inaccuracy of their use by analysing the colligation of the modals with bare infinitive, passive infinitive, progressive infinitive, perfect infinitive and perfect passive infinitive. Abdul Kader, Begi and Vasegi (2013) reported very

limited range of modals used by both levels of students. Only modals of ability *can* and *could* were most frequently employed by the students, which the researchers attributed to the argumentative nature of the student essays. There was also overuse of the modals *can* and *will*, which the author suggested was a result of transfer from the Malay equivalents '*boleh*' and '*akan*'. The study also concluded that in general students were able to use modals correctly in terms of forms and functions. Mukundan et al. (2013) also reported frequent use of *can* and *will* and added that 93% of the modals were syntactically accurate. However, the study did not present the statistical details of the accurate use according to the different colligations, which could reveal more information on which combinations the students mostly used and how they were used.

The data from MCSAW were also analysed for the use articles (*a, an, the*). Abdul Rahim, Abdul Rahim and Chia (2013) investigated the distribution patterns of these articles and their colligation patterns in the argumentative essays compiled for MCSAW. WordSmith Tools was employed to analyse the raw data from the corpus. The findings revealed that articles made up 18% of the entire corpus and definite article *the* recorded the highest percentage of use (74.7%). Article *the* also recorded the two highest colligation patterns *the + singular noun* and *the + plural noun*. No explanation, however, was rendered on the patterns that availed from the corpus, leaving questions as to why learners prefer these patterns or what possible factors could have determined the learners' choice.

Collocation studies are very rare in Malaysia, only a handful of studies have been published on the area so far and among them include Hong, Rahim, Hua and Salehuddin (2011), Abdullah and Noor (2013) and Joharry (2013). Hong et al. (2011) and Abdullah and Noor (2013) investigated the use of verb-noun collocations among Malay ESL learners in Malaysia. Hong et al. (2011) made use of EMAS corpus and BNC was used as the reference corpus. The focus of the study was on verb-noun collocations

errors, which the researchers found to stem mostly from wrong choice of the verbs or wrong form of the nouns. Abdullah and Noor (2013) analysed the similarities and differences in the use verb-noun collocations between Malay learners of English with native speaker learners. For the study, the researchers made use of a specifically compiled learner corpus; Written English Corpus of Malay ESL learners (WECMEL) consisting of argumentative essays written by pre-degree students from a public university in Malaysia. Contrastive analysis of WECMEL and LOCNESS (the native speaker learner component of ICLE) was administered. Unlike most LCR in Malaysia, Abdullah and Noor (2013) employed both quantitative and qualitative approaches to data analysis. Three (3) lexical verbs were recorded to occur most commonly in both learner corpora namely *reduce*, *make* and *take*. Malaysian learners were also found to prefer VO (verb-object) collocation combination similar to the native speaker learners. Nevertheless, the qualitative analysis revealed that despite these similarities the ESL learners had the tendency to use the lexical verbs differently than the native speaker learners. The study highlights the importance of supplementing quantitative analysis with a qualitative one as the latter would unearth the salient features of the learner language, which are often concealed from the aggregated data.

Another collocation study by Joharry (2013) focused specifically on the collocation and semantic prosody of the lemma *CAUSE*. The study aimed to shed lights on Malaysian learners' awareness of the negative prosodic feature of *CAUSE*. The corpus used was made up of ten (10) argumentative essays written by written a final assessments for third year undergraduate from several private universities in Malaysia. AntConc was employed for analysis of raw data. The research concluded that the collocational behaviour of the lemma *CAUSE* by Malaysian ESL learners was more inclined to negative evaluation consistent with that reported in previous studies and that semantic prosody was present in the Malaysian ESL learner writings.

LCR in Malaysia has also slowly begun to tackle discoursal aspects of learner language. The more recent study published on this aspect was by Aziz, Jin and Nordin (2016) on the use of metadiscourse among male and female ESL learners to construct their gender identities. The learner corpus utilised for the study was a compilation of argumentative essays written by third semester students from two higher institutions in Malaysia. The study compared and contrasted the employment of interactional resources (Hyland & Tse, 2004) between male and female learners and examined the extent the use of the metadiscourse reflects the articulation and construction of learners' gender identity. The main findings of the study suggest far greater similarities in the use of key interactional metadiscourse resource across gender. However, there were slight differences in stance making between genders and the way writers position themselves in the reader-writer interaction. Female learners were more assertive in expressing their stance, but at the same time would often include themselves in the reader-writer interaction by the use of inclusive *we*. In contrast, male learners preferred the more subtle way of stance making, but exhibited more dominance in the reader-writer interaction with the preference of pronoun *you*. This study marks the positive development of LCR in Malaysia, which sees the shift from lexicogrammatical, collocations studies to discourse. The study also made use of qualitative data to supplement the findings obtained from the quantitative analysis. The qualitative data provided the samples of how learners actually use the interactional resources in their essays.

Most recently, Mohd Don and Sriniwass (2017) conducted a study on the discoursal aspect of learner corpus using MACLE data. They analysed the use of conjunctive adjuncts in the subset of 54 argumentative essays written by Law students. The study identified 307 conjunctive adjuncts, which were grouped under three main categories namely (i) Elaboration, (ii) Extension and (iii) Enhancement. The learner data in this

study were not compared to that of the native speaker and the learner data were analysed in their own right, bringing what the learners can do and cannot do with conjunctive adjuncts in their essays. According to the researchers there was a clear mismatch between what students can do with what they are expected or required to do. Similar to Aziz, Jin and Nordin (2016), this study complemented the quantitative data obtained from the corpus analysis with qualitative data to support and explain the quantitative findings.

Another area of discourse investigated most recently was by Muthusamy and Farashaiyan (2017) on the patterns of compliments in writings of Malay ESL learners and English native speaker learners. The Malay learner data were obtained from the Malaysian Learner English Corpus, while the American university student component of LOCNESS was used as the reference corpus. The findings from the quantitative analyses revealed that Malay ESL learners constructed their compliments using action verbs followed by infinitives rather than with stative verbs, which is a norm among the native speakers. Even though the study managed to capture the differences in the patterns of compliments used by the learners through the aggregate data, it lacked exemplification of how the patterns were actually realised in both the ESL and the NS learner essays.

LCR in Malaysia has also begun to apply corpus-based methodology on second language rhetorical analysis. One such study was conducted by Botley (2014), who used a corpus of ten essays extracted from CALES to analyse the argument structure employed by Malaysian undergraduates in their written argument. The study concluded that even though learners were capable of constructing basic arguments, they appeared to use non-preferred structures, which bore very little similarities to the structures they were exposed to in their writing classes. There was also clear lack of depth and complexity in the learners' arguments, which the researcher attributed to low

proficiency in English. As the first study to apply corpus-based rhetorical analysis on second language writings, this study contributes tremendously to the LCR in Malaysia. It has expanded LCR in Malaysia and the same time paved the path to other similar studies to be conducted in the near future. Its contribution could have been more significant if the corpus-based methodology undertaken was detailed out. The study involved identification of argument structure that would require for the data to undergo an annotation process with the use of for instance a parsing tool. Since the annotation system and annotation tool were not specified and explained, replication study involving especially larger learner corpora would be difficult.

Although LCR has established itself as a vibrant research strand worldwide (Gilquin & Granger, 2015), its progress in Malaysia is rather slow-paced. Review of the LCR in Malaysia brings into focus several major aspects of the research strand needing much attention including firstly data accessibility. Most studies on learner corpora in Malaysia are limited to only several learner corpora, namely EMAS corpus (e.g. Arshad & Hawanum, 2010; Hong et al., 2011; Kamarudin, 2013; Zarifi & Mukundan, 2014) and MCSAW (e.g. Abdul Kader, Begi & Vasegi, 2013; Abdul Rahim, Abdul Rahim & Chia, 2013; Loke, Ali & Zulkifli Anthony, 2013; Mukundan et al., 2013), while other corpora like MACLE and CALES recorded very few research. This is because all of these learner corpora (MACLE, EMAS, CALES and MCSAW) are only available upon request to the respective developers making access to them relatively more difficult than a corpus that is available online. There is a need to increase the accessibility of these corpora in order to encourage more LCR to be conducted using the Malaysian ESL data. Such studies will not only contribute to LCR, but also to the teaching and learning of English in Malaysia.

Another issue concerning corpus data in some studies reviewed is the size of corpora examined. Studies conducted for example by Kanestion et al. (2016) involved analysis

of only 12 argumentative essays, while Joharry (2013) and Botley (2014) analysed only ten (10) essays each. Corpus-based approach is valued for its rigor in its design, data collection and analysis (Connor, 2004). It is a powerful methodology allowing concurrent analysis of sizeable corpora to be administered with speed and ease with the aid of text retrieval software tools. Investigation with extremely small size corpus would certainly affects the breath and the depth of the investigation, as many aspects of the learner language might not be available for analysis. More importantly the quantitative patterns drawn from the data would be too small to be generalised to the learner language the corpus aims to represent.

Studies on the grammatical aspects reviewed in this section also vary in terms of their methodological construct. Except for Kamarudin (2013), Hong et al. (2011), Abdullah and Noor (2013) and Muthusamy and Farashaiyan (2017) that adopted Granger's (2002) Contrastive Interlanguage Analysis, comparing learner language to that of the NS language to establish the underuse and overuse features, most of the studies did not make use of a NS corpus as a reference corpus. The studies concentrated mainly on describing the Malaysian ESL learner language in its own right (e.g. Abdul Kader, Begi & Vasegi, 2013; Abdul Rahim, Abdul Rahim & Chia, 2013; Arjan, Abdullah & Roslim, 2013; Kanestion et al., 2016; Loke, Ali & Zulkifli Anthony, 2013; Mukundan et al., 2013; Zarifi & Mukundan, 2014). Mohd Don and Sriniwass (2017) pointed out that ESL learner language does not necessarily need to be compared to the NS language and should not be constricted to the underuse, overuse and misuse analysis framework. The comparison would only quantify the linguistic form as being more or less in the learner corpus in reference to the reference corpus, when the learner corpus can be analysed in its own right to uncover aspects unique only to it.

According to Gilquin and Granger (2015) one of the major weaknesses of LCR is its exclusive dependence on aggregate data. It is criticised for being "relatively weak in

explanation" and they "often remain rather descriptive, documenting differences between learner and native language rather than attempting to explain them" (Myles, 2005, p. 380). Similar situation is also evident in the corpus-based studies in Malaysia. Most of the studies involved mainly quantification of a linguistic form under study with no focus on the how the form is used in context (e.g. Abdul Rahim, Abdul Rahim & Chia, 2013; Joharry, 2013; Loke, Ali & Zulkifli Anthony, 2103; Muthusamy & Farashaiyan, 2017). Nevertheless, there have been moves to integrate qualitative method to corpus data analysis (e.g. Abdullah & Noor, 2013; Aziz, Jin & Nordin, 2016; Mohd Don & Sriniwass, 2017). Realising the importance of supplementing quantitative findings with qualitative ones, the investigation on the use of *BE* in the current study employs both types of data analysis. The distributional patterns of the forms and functions of *BE* will be obtained from the quantitative analysis, while the qualitative findings provide the explanation on how *BE* is actually used in the learner essays.

There is also the need to expand LCR to include more investigations on annotated data as the majority of learner corpus studies are based on raw data (Gilquin & Granger, 2015). The studies adopting CEA approach reviewed in this section were all conducted on raw data (e.g. Arshad & Hawanum, 2010; Jishvithaa, Tabitha & Kalajahi, 2013; Manokaran, Ramalingam & Adriana, 2013). Learner errors were not annotated and no annotation system was adopted from past studies or developed for the purpose of these studies. Studies on learner errors would entail for the data to be annotated (e.g. Granger, 2003; Thewissen, 2013). The lack of systematic tagging system raises concerns regarding the findings, as raw data would not be able to reveal hidden aspects that could only be visible with annotated data for example the syntactic environments in which errors are more prone to. Applying annotation on the data would also enable speedy and easy retrievals of the errors with the use of any standard text retrieval software tools and more importantly allowing all analyses to be conducted

automatically (Granger, 2003). More importantly annotating the errors would enable researchers to systematise and profile the learner errors, which can be manipulated in comparative studies between/within learner corpora. According to Gilquin and Granger (2015), annotated data would also enable the linguistic analysis to be further expanded in the future to include aspects that were not manipulated in previous studies. The present study intends to conduct a corpus-based investigation on annotated learner corpus. A manual tagging system will be developed for the purpose of annotating grammatical and ungrammatical uses of *BE* and the constituents before and after *BE*. The system will be the first to be developed in LCR in Malaysia and can be used as a reference for similar studies in the future.

Another aspect concerning data annotation that demands attention is the use of automatic POS taggers (most notably CLAWS) in tagging learner corpora. Automatic POS taggers like CLAWS are trained on the NS language, they are not designed to detect and tag learner errors. Using automatic POS tagger on learner data that contain learner errors could significantly affect tagging accuracy. van Rooy and Schäfer (2002) compared the performance of three automatic taggers, namely CLAWS, TOSCA-ICLE, and the Brill tagger on the Tswana Learner English Corpus (TLEC) and found that accuracy was affected by errors in spelling, lexical choice, verb conjugation, clause type, the use of the infinitive and omissions. In view of the possible threats of learner errors on tagging accuracy of automatic POS tagger, the present study devices and develops a manual tagging system to manually tag the grammatical and ungrammatical *BE* and the constituents before and after *BE* in the study.

In addition, detailed and comprehensive investigation on specific aspect of grammar especially involving annotated learner data has also been under investigated in Malaysia. Even though *BE* has been investigated in previous studies (Arshad & Hawanum, 2010; Jishvithaa, Tabitha & Kalajahi, 2013; Manokaran, Ramalingam and

Adriana, 2013), the focus was only on the errors on auxiliary *BE*. Other functions of the verb (e.g. copula, negative operator, interrogative operator) have not been investigated thus far, and very little is known about what learners can and cannot do with *BE*. In order to grasp the extent of learner acquisition of *BE*, there is the need to analyse not only the errors, but also the correct use of the verb. The present study intends to conduct a comprehensive analysis of *BE* that would unveil the patterns of the grammatical and ungrammatical uses of *BE* and whether the uses are influenced by the syntactic environments.

2.1.3 Approaches to Learner Corpora Research

Analysis of learner corpora can fall into two methodological approaches; Contrastive Interlanguage Analysis and Computer-aided Error Analysis (Granger, 2002). The subsequent sections discuss these approaches further.

2.1.3.1 Contrastive Interlanguage Analysis

Contrastive Interlanguage Analysis involves either NS/NNS or NNS/NNS comparison as shown in Figure 2.1. NS/NNS comparison can yield valuable information on the range of features unique to the NNS, not only the deviant forms and structures but also the patterns of use. By comparing the patterns of use with that of the native speakers, linguists would be able to determine the degree of differences (or similarities), thus, enable work towards closing the gap to be carried out.

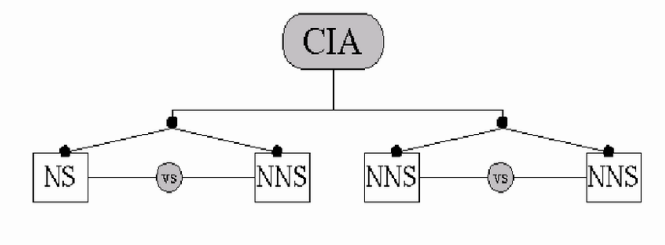


Figure 2.1: Computer Interlanguage Analysis

Granger (2002, p.12)

Many corpus-based studies were conducted using the NS/NNS comparison, one such study was by Aijmer (2002) comparing the range and frequency of English modal words in the written language of native speaker learners (LOCNESS) with L2 learner writers. The findings revealed among others, that NNS learners' overused of the global modal devices, which was the result of them adopting speech-like style in their writing. Their usage patterns differed greatly with the NS data from the LOCNESS corpus. Other learner corpora research adopting NS/NNS comparison include Granger and Rayson (1998), Hong et al. (2011), Abdullah and Noor (2013), Kamarudin (2013), Zarifi and Mukundan (2014) and Gilquin and Granger (2015) (refer to Section 2.1.2.1).

The second type of comparison (i.e. NNS/NNS) allows researchers to study either the developmental or L1 dependent features between learners of different L1s, or within the same L1 group with different proficiency levels. Granger and Tyson (1996) conducted a NNS/NNS comparison in their investigation on connectors by learners of different L1s. The findings showed how certain features like the overuse of sentence-initial connectors to belong to developmental type feature, as they were found to be consistent in at least three of the learner populations (French, Dutch and Chinese), and variation in the use of individual connectors between learner groups exhibited characteristics of L1 influence (Granger, 2002). The findings of this study provide support to the importance of conducting NNS/NNS comparison as it can help researchers to differentiate between transfer and developmental features. Gilquin and Granger (2015) stressed the importance of selecting comparable corpora preferably those built with the same design criteria (e.g. ICLE) for the purpose of NNS/NNS comparison. Adel (2008) for instance found that factors such as different composition of the L1 sub-corpora and proportion of timed versus untimed essays in the learner sub-corpora resulted in the differences in the use of involvement features in argumentative essays of these learners. Gilquin and Granger (2015) propose researchers to take into account variables other than

interlingual transfer such as learner's proficiency, age, gender, and knowledge of other languages or access to reference tool in their NNS/NNS comparison.

The current study adopts the NS/NNS comparison, with the corpus data of the NS variety obtained from Louvain Corpus of Native English Essays (LOCNESS), while the non-native speaker data from the Malaysian Corpus of Learner English (MACLE). Nevertheless, the study does not intend to highlight the underrepresentation and overpresentation features in the NNS variety as prescribed by Granger (2002). Instead the comparison is conducted to attain the similarities and differences in the use of *BE* by the NS and NNS learners.

2.1.3.2 Computer-aided Error Analysis

Computer-aided error analysis (CEA) devotes entirely on the analysing errors in the learner corpora. As evident in the term, the use of specialised computer tagging software is required for this type of analysis. This sets it apart from the traditional Error Analysis (EA) framework; it allows for vast quantity of data to be analysed, covering more error types in lesser time as it might take the traditional EA.

CEA analysis can be carried out using two methods, the first is by selecting error prone linguistic items and scrutinise the corpus for the pre-determined error types. This can be done with the help of standard text retrieval software tools such as WordSmith Tools (Scott, 2017). The analysis process is rather straight forward; it involves scanning the corpus to retrieve all the instances of the pre-determined error types. One most important advantage of this method is the use of a text retrieval software results in extremely fast error retrieval. Nevertheless, the method has its limitations. The search and analysis of data are restricted to only areas or items that are considered problematic.

The second method of CEA involves a more laborious scrutiny of the corpus. It entails the researcher to devise "a standardised system of error tags and tagging all the errors in

a learner corpus or, at least, all the errors in a particular category..." (Granger, 2002, p. 14). This method is undoubtedly more time-consuming, however, according to Granger (2002) it is much more powerful as it can reveal learner difficulties that researchers might not be aware of. This method requires for laborious and intensive error-tagging process. The tagging of errors can be made easier with the help of an error editor for instance the *Universite Catholique de Louvain Error Editor* (UCLEE), which is available at Louvain University (Dagneaux, Denness & Granger, 1998). Once the error-tagging work is completed, the errors can be retrieved easily using standard text retrieval software. One advantage of this method is once the corpus is fully error-tagged, it can be used for a wide range of language investigations as pointed out by Granger (2002) "the possible applications that can be derived from it is absolutely huge" (p.14).

Dagneaux et al. (1998) introduced the technique of CEA and highlighted the potential of CEA over traditional EA. With the help of an MS Window error editor (UCLEE), the researchers had error tagged a 150 000-word French learner corpus. The corpus consists of argumentative essays written by advanced and intermediate learners of English. The researchers demonstrated that error-tagged corpus made it possible for the learner population to be categorised in terms of the major error categories. More importantly the study has managed to counter the limitations of traditional EA. CEA is able to (i) generate results for not only learner errors but also non-errors and (ii) provide more dynamic picture of L2 learning by comparing the language of L2 learners at two different stages of their curriculum, something that would not be possible with traditional EA.

CAE analysis is also adopted in the current study since it also involves the analysis of the ungrammatical use of *BE*. Since CAE entails for the learner errors to be annotated, the learner errors in MACLE is manually annotation by using a set of standardised

system of error tags devised especially for this purpose. A manual tagging tool (Malaysian Linguistic Tagging Tool), which was especially developed for this purpose, is employed for the manual tagging task.

2.2 Application of Corpora in Language Teaching and Learning

In providing the answer to the "so what" question, this study provides suggestions on how corpus consultation can be applied in the teaching of *BE*. This section provides an overview of corpus consultation, a review of previous studies that have incorporated corpora in the L2 classrooms and finally the learners' attitude towards the incorporation of corpora.

2.2.1 Overview of Corpus Consultation

The integration of corpus in the context of L2 learning and teaching has become increasingly appealing in the recent years (Gaskel & Cobb, 2004; Leel, 2011; Phoocharoensil, 2012; Shi, 2017; Vyatkina, 2017; Yoon & Hirvela, 2004; Yoon, 2008, 2011; Yunus & Awab, 2012, 2014). The ability to simultaneously integrate language skills such as vocabulary, grammar, writing and reading skills makes corpus consultation an attractive complement to the traditional method of language teaching and learning. The corpora, which are collection of massive language database from multiple resources, offer learners with a rich exposure to the genuine language (Thurstun & Candlin, 1998; Vyatkina, 2017; Yoon & Hirvela, 2004; Yunus & Awab, 2012, 2014), the kind that they will most likely encounter and eventually use outside the classroom. The exposure can enrich learners' understanding and repertoire of the target language (Yoon & Hirvela, 2004), hence contribute to the overall growth in it.

Corpus consultation literacy (O'Sullivan, 2007) in the language teaching and learning is valued for its ability to promote inductive learning, where learners take central stage in their own language learning process (Johns, 1991; Lee & Swales, 2006; O'Sullivan,

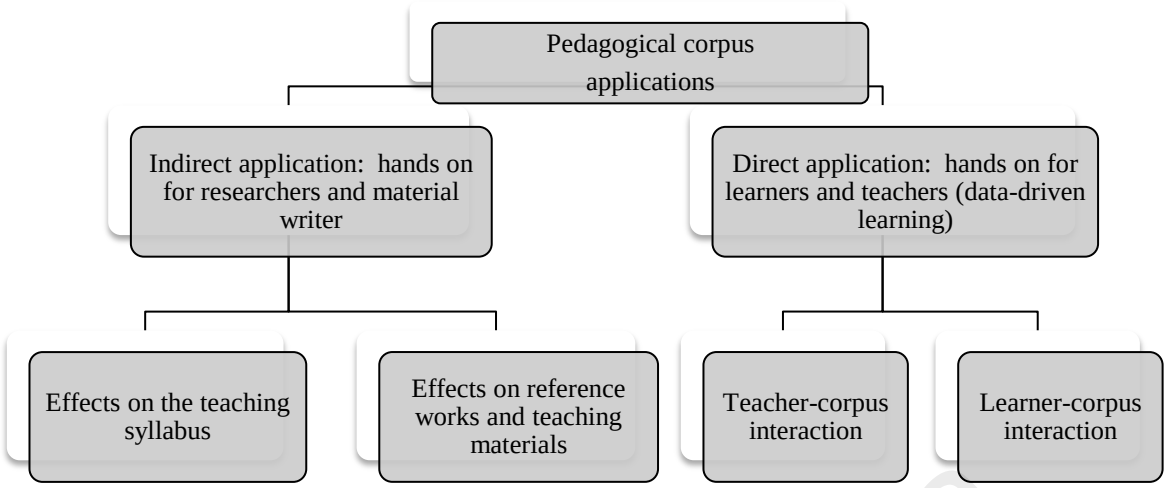
2007; Todd, 2001). In his pioneering work on data-driven learning (DDL), Johns (1991) introduced the learners as language researchers, who perform the role of research workers or according to Bernardini (2004) language explorers/travelers, whose task is to explore the language. The approach requires for the learners to be directly confronted with the language, analysing it and deriving to the language patterns. This exploratory nature of DDL can be highly "motivating and highly experiential" (Kettemann, 1995, p. 10) for the learners and can be much more interesting and rewarding than being taught about language (Phoocharoensil, 2012; Shi, 2017).

Concordancing is also believed to be a powerful tool to develop learners' cognitive processes. Analysing concordances involves various forms of cognitive skills which include "predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying" (O'Sullivan, 2007, p. 277). O'Sullivan (2007) then added that these activities not only have the potential to increase learners' mental activity, but also help them to develop their learning and cognitive processes. Boulton (2010, 2011) pointed out that the inductive and discovery learning involved in corpus consultation stimulate learners' deep processing which may lead to more effective learning. Corpus consultation also has the potential to develop learners' analytical skills (Barbadi & Khajavi, 2017; Shi, 2017; Yunus & Awab, 2012, 2014), problem-solving abilities (Barbadi & Khajavi, 2017; Johns, 1991; Johns, 1997; Kennedy & Miceli, 2001) and increase their observation, attention, and capacity for reflection (O'Sullivan, 2007). These thinking and learning skills and strategies not only increase learners' cognitive abilities, but also "can be used to resolve many other language problems" (Stevens, 1995, p. 2).

The computer-based nature of corpus consultation has also added value to the approach. The use of computer gives the L2 learners greater exposure to the target language and

allows for bigger opportunities for interaction with it (Yoon & Hirvela, 2004). The abundance of computer-based artefacts such as the internet and hypertext and the availability of online corpora (e.g. Bank of English sampler, Collins COBULID sampler and Michigan Corpus of Academic Spoken English) allow for limitless access to the target language texts (Conrad, 2000; Sun, 2000). It also allows for more freedom to the L2 learners, learners not only gain access to endless texts to choose from but also given the upper hand to select texts that would have "the greatest linguistic value" relative to their study needs and requirements (Yoon & Hirvela, 2004, p. 261). By comparison the opportunities to interact and work with texts specified to the learners' discoursal domains can be more linguistically rewarding than working with the general texts available in the textbooks. Most importantly direct access to the vast arrays of language resources, which are readily available round the clock, encourages the learners to use these resources outside the classrooms, thus provides ample opportunity for them to exercise autonomy in their own learning (Yoon & Hirvela, 2004).

Romer (2011) in his overview of pedagogical corpus applications made a distinction between direct and indirect applications of corpus in the process of language teaching and learning. According to the researcher indirect corpus applications can help teachers to decide *what* and *when* to teach and they can have profound effects on the teaching syllabus and the design of the teaching materials, while direct applications affect *how* something is to be taught or learned. Direct applications involve active participation of the learners and teachers in the process of working with the corpora and concordances. Figure 2.2 below summarises the pedagogical applications of corpora:



Romer (2011, p. 207)

Figure 2.2: Pedagogical Applications of Corpora

2.2.2 Previous Studies on Corpus Consultation in Language Pedagogy

Classroom-based research on learning outcomes of DDL has been carried out in several areas of language learning. Much of the research has been centered in the field of vocabulary learning (e.g. Barabadi & Khajavi, 2017; Boulton, 2010, 2011; Cobb, 2007; Guan, 2013; Miceli & Kennedy, 2002; Shi, 2017). Shi (2017) for instance reported statistically significant relationship between corpora incorporation and learners' knowledge in vocabulary. Learners trained with and exposed to corpora scored better in the post-tests than learners who underwent conventional vocabulary teaching method. The researcher stressed that pedagogical application of corpora with adequate instruction can be effective in improving learners' vocabulary. Barabadi & Khajavi (2017) echoed similar conclusion, stressing that better performance of the experimental group in their investigation can be attributed to the active role learners played in the learning process through self-discovery, inductive and bottom-up processes.

Positive results were also reported for incorporation of corpora in teaching grammar (Chambers & O'Sullivan, 2004; Gaskel & Cobb, 2004; Leel, 2011; O'Sullivan & Chambers, 2006; Phoocharoensil, 2012; Todd, 2001; Vannestal & Lindquist, 2007;

Yoon, 2008; Yoon & Hirvela, 2004). Many of these studies have measured learners' attitudes and experiences with DDL and grammar instruction, and some have also measured learning outcomes and found benefits to using DDL in grammar instruction. Leel (2011) for instance with the utilisation of *CONTEXTS* program showed how the concordances of grammatical items could be used to help learners learn prepositions. Learners were presented with samples of concordances and assigned exercises of selected prepositions prepared with the use of *CONTEXTS* program. The implementation of DDL according to Leel (2011) provided an alternative for the highly prescriptive language learning, which was the normal practice in the Taiwan L2 classrooms. Through this pilot analysis the researcher argued that DDL has a great potential to prepare students for examinations as well as for general English acquisition.

Integration of corpora has also been found to increase ESL learners' collocational competence. Yunus and Awab (2012, 2014) reported that ESL learners significantly benefitted from the application of both paper-based concordance materials and computer-based DDL in enhancing the knowledge of collocations of prepositions. The findings from the experimental study indicated that DDL group performed significantly better in sentence-completion, error-identification, error-correction and semantic-function tasks compared to CA (conventional approach) group. The same success was also reported in Vyatkina's (2017) study on German verb-preposition collocation among North-American college students. Similar to Yunus and Awab (2012, 2014), paper-based and computer-based DDL activities were used to explore learners' ability to perform in the grammatical tasks prepared. Vyatkina (2017) concluded that DDL intervention was effective in increasing learners' overall proficiency in verb-preposition collocation. These studies highlighted the effectives of both hands-on and hands-off DDL approaches in developing L2 grammatical knowledge.

Corpora intervention has also made considerable mark in improving general writing

skills (Chambers & O'Sullivan, 2004; Gaskell & Cobb, 2004; Kennedy & Miceli, 2001; Kotamjani, Razavi & Hussin, 2017; Miceli & Kennedy, 2002; O'Sullivan & Chambers, 2006; Yoon, 2008; Yoon & Hirvela, 2004). Miceli and Kennedy (2002) and Kennedy and Miceli (2001) observed how integration of corpus was successful in developing less advanced learners' writing skills in Italian by means of corpus-based error correction and content and vocabulary enriching activities. Chambers and O'Sullivan (2004) in their research involving advanced learners of French at the University of Limerick concluded that learners in general were able to make positive changes to grammatical, misspellings, lexico-grammatical patterning and capitalisation errors after the implementation of corpus intervention. Kotamjani, Razavi and Hussin (2017) also reported improved skills in EFL academic writing among university undergraduates after they were taught and trained to use corpus as a reference. Learners' ability to proofread and edit the surface level of their writing had improved significantly and this consequently builds their confidence in writing.

From the studies reviewed, it is clear that direct corpus consultation has potential in the teaching and learning of language. Learners' responses and the positive results in error correction and deducing patterns activities have given support to such claim. The findings from the literature reviewed in this section shall be referred to as a guide to the corpus consultation model proposed for the teaching of *BE* in the current study.

2.2.3 Learners' Attitude towards Corpus Consultation

Besides the focus on the effects of corpus consultation on the teaching and learning of L2, researchers have also begun to turn their attention to learners' attitude towards the incorporation of corpora in their language classrooms. Sun (2000) explored a group of Taiwanese ESL learners' perception towards web-based concordancing. The findings revealed that the learners had positive attitude towards the web-based concordancing

activities. They felt mostly positive about the opportunity to gain access to authentic language use and they also viewed the approach as most helpful in improving their reading comprehension skills and in acquiring knowledge of actual use of words and phrases.

Positive results were also obtained in the incorporation of corpus as a writing tool. Yoon and Hirvela (2004) and Yoon (2008) reported learners had positive feeling about the use of corpus in improving their writing. They believed that corpus had improved their general writing skills. In general learners valued the usefulness of the corpus in providing textual help in the writing. Incorporation of corpora was also positively received in grammar classrooms. Phoocharoensil (2012) investigated Thai EFL students' attitudes towards corpus-based grammar lessons on *if*-clause and relative clause *who* and *whom*. The students perceived the incorporation of corpus-based activities to be beneficial in learning grammar. Most of them preferred learning grammar through concordance lines than other learning methods. In addition, they also valued the experience of rule discoveries, which gave them more sense of accomplishment and satisfaction. Additionally, Charles (2007) in her study investigating how corpus-based activities can be used to complement discourse analysis in the process of raising learners' awareness of the rhetorical functions of selected thesis components also found that learners were generally positive towards this approach.

Nonetheless, negative responses to corpora consultation have also been attested in previous studies. Vannestal and Lindquist (2007) in investigating how learners' attitudes about grammar were affected by the introduction of corpus-based activities as complements to the ordinary grammar textbooks and exercises, found that learners in general were less than enthusiastic with this approach. The researchers concluded that the corpus activities were not able to foster positive changes in the learners' attitude towards learning grammar and their expectation regarding proficiency, understanding

and explaining grammar rules were not fulfilled as hypothesised by the researchers. The researcher identified several factors influencing this result namely, difficulty accessing the corpus, lack of training, and difficulty adapting to the corpus-based learning approach.

As attested by the studies reviewed, with proper implementation and sufficient training corpora intervention can be effective in helping learners overcome various linguistic problems. In addition, its implementation permits more learner autonomy, as learners can work independently to discover the language rules and patterns, which is relatively more interesting and rewarding than being taught about them.

2.2.4 Summary of Studies on Corpus Consultation

Corpora in the language pedagogy can function as either a research tool or a reference tool (Yoon, 2011). A corpus is considered a research tool when it is used by the learners to conduct linguistic investigation to infer language patterns or rules. In this context the learners perform the role of linguistic detectives (Johns, 1991), who actively research or investigate the language to infer language patterns available from the corpus. Studies done by Todd (2001), Cheng et al. (2003), Vannestal and Lindquist (2007) and Farr (2008) are some of the studies in which the corpora were used as research tools. Nevertheless, this approach to language learning can pose serious challenge to learners who are not accustomed to the inductive approach to language learning (Vannestal & Lindquist, 2007; Yoon & Hirvela, 2004). Not many learners possess genuine interest in building and testing language hypotheses or researching a language (Kennedy & Miceli, 2001; Miceli & Kennedy, 2002). They may find the tasks of conducting concordance searches and interpreting the search results as difficult and challenging (Farr, 2008; Vannestal & Lindquist, 2007; Yoon, 2008; Yoon & Hirvela, 2004). In order to conduct successful searches and to accurately infer patterns learners

need to possess a sound knowledge of the target language (Boulton, 2010, 2011; Chambers & O'Sullivan, 2004; Cheng et al., 2003; Gaskell & Cobb 2004). For that reason this approach is argued to benefit more advanced learners (Boulton, 2010, 2011; Chambers & O'Sullivan, 2004; Granath, 2009; O'Sullivan & Chambers, 2006).

A corpus is categorised as a reference tool when it is mainly used as a "linguistic reference" or a reference tool "to solve writing and language problems" in the process of revising and improving learners' writings (Yoon, 2011, p. 134). This approach has received positive responses from the learners as it is less demanding and involves mostly "observe and borrow" learning activities instead of "observe and derive rules" activities (Miceli & Kennedy, 2002). Learners would be engaged in observing the language in the process of improving a small part of their writing. This approach is more suitable for beginners to corpus consultation, as they would not be subjected to the linguistic burden and "cognitive burden" (Boulton, 2010) that would surface in a full scale linguistic investigation. Corpora have proven to be an important reference tool in the teaching of writing and grammar (Chambers & O'Sullivan, 2004; Gaskell & Cobb, 2004; Kennedy & Miceli, 2001; Kotamjani et al., 2017; Leel, 2011; Miceli & Kennedy, 2002; O'Sullivan & Chambers, 2006; Phoocharoensil, 2012; Yoon, 2008, 2011; Yoon & Hirvela, 2004) mostly means of error correction. Phoocharoensil (2012) for instance reported that Thai students, who were exposed to corpus-based error correction activities perceived the incorporation of corpus to be beneficial in learning grammar. Most of the students preferred learning grammar through concordance lines than other learning methods. In addition, they also valued the experience of rule discoveries, which gave them more sense of accomplishment and satisfaction.

Another important consideration in the use of corpora in language teaching is the selection and size of corpora. Corpora used in teaching can be divided to two namely, specialised corpora and general corpora. Specialised corpora are strongly associated

with the teaching of Language for Specific Purposes (LSP), where the use of custom-built corpora is motivated by the need to cater to the linguistic requirements and interests of specialised linguistic discourses. As an example, the use of Contemporary Written Italian corpus (CWIC) for the teaching of writing skills in Italian to less advanced learners of Italian (Kennedy & Miceli, 2001; Miceli & Kennedy, 2002). Specialised corpora are generally smaller in comparison to general corpora.

General corpora are bigger and are commonly used to teach general aspects of language such as phrasal verbs or *BE*. General corpora are better suited for the teaching of more general aspects of language that learners regardless of their majors or disciplines find difficult or problematic (Yoon, 2011). The larger size of general corpora (e.g. BNC-100 million words, Collins COBUILD-500 million words) adds to the advantage of using them. They are very useful in providing countless samples of language usages from many different registers, disciplines or contexts. The corpora provide the learners greater opportunities to interact with the authentic language used in the various contexts, disciplines or registers. Some of them can be accessed free of charge online (e.g. British National Corpus, Collins COBUILD and Corpus of Contemporary American English) and most importantly some are already equipped with inbuilt concordancers, which allow for simple concordancing to be administered (e.g. British National Corpus, Collins COBUILD and International Corpus of Learner English). This is an important criterion to be considered as not all teaching institutions or schools have the resources and budget to acquire licenses for commercially built concordancers such as WordSmith Tools (Scott, 2017).

In terms of the actual corpus-based activities conducted, the literature has narrowed them down into two major activities: self-correction of errors and patterns deducing activities. Self-correction is considered an important process in language learning and it is "viewed as a global goal of language learning" (Todd, 2001) since the ultimate goal

of language learning is for the learners to be able to initiate self-repair (Allwright & Bailey, 1991). The results from past studies (e.g. Todd, 2001; Miceli & Kennedy, 2002; Kennedy & Miceli, 2001; Gaskell & Cobb, 2004; Chambers & O'Sullivan, 2004; O'Sullivan & Chambers, 2006) indicate that corpus consultation is an effective approach to word and sentence-level error correction. The overall percentages for accurate error corrections were high in all the studies except for Gaskell and Cobb (2004), which obtained mixed results. Based on the positive learner feedbacks and the overall results of accurately corrected errors in the other studies reviewed, corpus consultation can be a more interesting and effective approach to error correction and at the same time has proven to be successful in training learners to become independent users of concordancers (Yoon, 2011).

Deducing patterns activities are useful to help learners overcome genuine language problems (Yoon & Hirvela, 2004; Yoon, 2008) and to enrich their knowledge in the target language especially in lexico-grammatical patterns, grammar, vocabulary and collocations (Chambers & O'Sullivan, 2004; Kennedy & Miceli, 2001; Miceli & Kennedy, 2002; Vannestal & Lindquist, 2007). However, they require learners to possess a fairly sound knowledge of the target language to be able to successfully deduce patterns or to even formulate search terms to begin with (Boulton, 2010; 2011). It is not surprising that the activities were more positively viewed by learners for error correction purposes and for helping them to check for grammaticality, vocabulary, gender and syntax (Chambers & O'Sullivan, 2004), where the corpora were consulted only for checking and confirming purposes much like a dictionary or a grammar reference (Yoon, 2008, 2011).

Another important aspect of corpus consultation is the learners' reaction and evaluation of the approach. Corpus consultation has received both positive and negative reactions from the learners. Across the studies, most of the learners reacted positively to the

corpus-based activities that they were engaged in and they valued that corpus consultation can provide:

i. authentic language,
ii. contexts where words structures are used,
iii. quick and easy referencing tool for checking and confirming,
iv. greater autonomy in learning, and
v. confidence in L2 writing.

Yoon (2011, p. 136)

The approach has also received several negative responses from the learners. One recurring issue is that learners find it time consuming to sort through concordance examples. They were overwhelmed with the sheer number of concordances and often frustrated for not finding the relevant examples or for generating too few or no concordances. As noted earlier for learners to be able to successfully use corpora, they need to possess a fairly sound knowledge in the target language or they would experience difficulties in forming search terms and interpreting concordances. For this reason corpus consultation is believed to benefit more advanced learners (Boulton, 2010, 2011; Johns, 1991; Turnbull & Burston, 1998).

Learners' difficulties were also found to stem from lack of training in handling and interpreting concordances and the technical aspect of concordancing. In some studies learners were not given adequate training prior to concordancing exercises and very limited guidance during the exercises (e.g. Vannestal & Lindquist, 2007). Proper training and guidance are most crucial in ensuring successful implementation of corpus consultation as pointed out by Miceli and Kennedy (2002) and Kennedy and Miceli (2001). The technical aspect of corpus consultation should also be stressed upon as learners no matter how advanced they are, might not have the computer skills to handle computer works and the relevant concordancing software (Yoon & Hirvela, 2004).

The negative responses can be summarised as the followings:

i. time consuming going through concordances and to find relevant ones,
ii. frustrating not to understand all concordance examples,
iii. hard to formulate proper search terms and interpret search results,
iv. frustrating to get too few or no concordances for search terms,
v. lack of training and guidance to conduct searchers and interpret results, and
vi. lack of training in operating the computer and concordancing software.

Yoon (2011, p. 136)

Based on previous studies, the success of any corpus consultation would be highly dependent on three focal elements, which include the linguistic features to be taught, the selection of the corpora and the learning activities to be carried out. In ensuring maximum success in its implementation, researchers and teachers also need to first acknowledge the difficulties faced by the learners, so that measures can be devised to overcome them. One focal element that could ensure successful integration of corpora is to provide learners' with adequate training in concordancing. Learners need gradual and guided training that can accommodate their different learning styles, experience and language proficiency levels (Yoon, 2011). Another aspect of training that is vital in the successful integration of corpus is to equip learners with general Information Technology (IT) literacy and the mastery of the general functions of the concordancing software. The lack of proficiency in IT related functions can also negatively affect learners' perception towards the approach (Yoon & Hirvela, 2004).

2.3 Previous Studies on *BE*

The literature reviewed in this section covers two main areas; (1) previous studies on *BE* in the first and second language acquisition research and (2) the position of *BE* in

the Malaysian English. Besides providing the empirical background of the study, the section also addresses the historical, theoretical as well as methodological issues relevant to the examination of *BE*.

2.3.1 *BE* in First Language Acquisition Research

Language acquisition studies have placed immense attention on how children acquire language in general and more specifically how they attain the syntactic knowledge of a language. In the interest of unraveling this mystery, many studies have been conducted investigating young children speech productions in the attempt to uncover the acquisition pattern evident in them. Researchers investigating this area are divided in their beliefs on what influences child acquisition of syntactic knowledge. The nativists (e.g. Becker, 2002; Chomsky, 1965; Schütze, 2004) assume the ability children have to acquire language lies solely on the innate linguistic faculty (Universal Grammar). They believe that children are born with an adult-like abstract representation of language functional system. In contrast, constructivists (e.g. Ambridge, Rowland, Theakston, & Tomasello, 2006; Theakston & Lieven, 2008; Theakston, Lieven, Pine, & Rowland, 2000; Theakston & Rowland, 2009; Wilson, 2003) argue against the innate hypothesis, forwarding instead a belief that language is acquired through exposure to the environmental input that contains rich data on language structure. Through research done on child L1 the two camps are able to provide evidence in support of their respective arguments. The following paragraphs present and discuss the findings from both camps, with focus on the acquisition of *BE*.

From the nativist camp, Becker (2002) using the data from the CHILDES database (MacWhinney, 2000), investigated the acquisition pattern of copula *BE* of three L1 English learners aged between 1;0 to 2;8. She found that the children frequently omitted *BE* with the omission higher with stage-level predicates than with individual-

level predicates, which led her to suggest that children's early production was sensitive to the types of predicates. The type of predicatives; nominal, adjectival and locative, according to Becker (2002) had a very strong influence on the omission of *BE*. The researcher observed that most nominals and some adjectivals were individual-level predicates and one characteristic of this type of predicates was they denote permanent properties (*he's a dog* and *she's happy*). Locatives and some other adjectivals, on the other hand, fall under stage-level predicates (*it's in the kitchen*), which always denote temporary properties (Becker, 2002). She suggested that stage-level predicates contain an additional aspectual projection, resulting in the clauses being realised as non-finite.

Early child language acquisition was also analysed in terms of the finiteness of *BE*. Schütze (2004) investigated omission rate of finite and non-finite *BE* among L1 English speaker in the CHILDES database (MacWhinney, 2000). The researcher uncovered an interesting aspect the acquisition of *BE* in particular the bare infinitive *BE* as in *Mary is gonna* ***be*** *nurse* was almost never dropped. Schütze (2004) only found very minimal instances of non-finite *BE* omission in the file of one of his subjects (Anne), which recorded a 21% rate of finite *BE* omissions compared to only 3% rate of non-finite *BE* omissions. Other files recorded zero omission of non-finite *BE*. Unlike finite *BE*, the root infinitive *BE* does not stand on the same relationship with Tense like the finite *BE*. Schütze (2004) argued that in the case of finite *BE* drop, the Tense is underspecified resulting in the condition where Verb Requirement no longer exists and the presence of *BE* is, therefore, not required. In a non-finite context, where the infinitival *BE* materialises alongside a modal or auxiliary, the Tense is realised on the modal or auxiliary, forcing the presence of *BE* when Tense imposes its Verb Requirement. Schütze (2004) argued that this was the reason why infinitival *BE* was almost never omitted in the child grammar he analysed.

In the attempt to shed more lights on the shape and principles governing early Inflectional Projection system (IP), Moscati (2006) conducted a cross-linguistic investigation on the pattern of copula *BE* omission in negative utterances of Italian and English child speakers. Analyses administered on the declaratives found a symmetrical pattern in the omission of copula *BE* in English and Italian. Becker (2002) recorded copular omission in the early grammar of English children ranging from 37% to 66%, which was not too far apart from those of Italian children with the omission rate between 49% and 81% (Moscati, 2006). The data from the negative contexts, however, did not mirror those of declaratives. Null copula *BE* was still attested in English utterances with an average of 21.4% of omission, but in Italian the omission rate was recorded at 0%. The asymmetry of null copula pattern in Italian and English negative constructions, according to Moscati (2006) can be attributed to the syntactic context of the two languages and the parametrical choice made at the age when the investigation took place. The data from this study present another aspect of null copula *BE* with respect to negation that brings into focus the syntactic relation to Tense, which influences the supply of copula *BE* in the early child grammar (Moscati, 2006).

Wilson (2003), presented a constructivist's view of the acquisition of English inflections (copula *BE*, auxiliary *BE* and 3sg present agreement) among English L1 children. The investigation involved analysis of longitudinal transcripts of five children ages between 1;6 to 3;5. The purpose of the investigation was to ascertain the condition of inflection emergence and to argue against The Agreement/Tense Omission Model (ATOM) (Pine et al., 2008; Rice, Wexler, & Hershberger, 1998) that posited the development of inflection as a unitary category. ATOM predicts a similar pattern of provision across different tense-marking morphemes resulting in the development for individual morphemes to be similar to one another. Wilson (2003) held the view of the constructivists, who argued that inflection develops in a "piecemeal fashion" and

heavily embedded in lexically specific constructions such as *He's/It's/I'm* (Pine et al., 2008; Rice et al., 1998).

Wilson (2003) observed that closed-class subjects like *he*, *she* and *I* occurred more frequently in the input which meant higher possibility of them occurring in specific construction with allomorph of *BE* as in *he's*, *you're* and *I'm*. He predicted since open-class subjects occurred less frequently in the input, there would be less chances of children producing construction like *pony's*. Wilson (2003) found that four out of five children produced copula and auxiliary *BE* significantly more with closed-class subjects than with open-class subjects. This finding further supported the constructivists' position that inflectional morphemes are acquired as chunks and to a large extent independent, thus, rejecting the nativists' notion that first language acquisition is determined by one underlying category.

Wilson's (2003) findings were later verified by a group of researchers Pine et al. (2008), who like Wilson aimed at providing empirical evidence on the constructivists' view on the provision of tense-marking morphemes in early child grammar. They conducted a longitudinal investigation on speech production of 11 English speaking children, who were between 1;10 to 3;0 of age at the time of the study. Replicating Wilson's (2003) analysis, the investigation was also set out to test the ATOM (Pine et al., 2008; Rice et al., 1998) prediction that there would be a similar provision pattern across different tense-marking morphemes in early child grammar. This was done by analysing the provision rates for 3sg present tense and first and 3sg forms of copula and auxiliary *BE* in the speech data obtained from Manchester Corpus.

The findings from the analyses showed considerable variation in the provision rates of the different morphemes, with 3sg copula *BE* provision higher than 3sg auxiliary *BE* and 3sg auxiliary *BE* higher than 3sg present tense inflection. Furthermore, there were

also variations in the provision of copula and auxiliary *BE* in the context of individual pronominal subjects *It, He* and *I.* These findings provide further evidence that children's knowledge of tense-marking morphemes relies on lexically-dependent constructions such as *'It's+NP'* and *'He's V-ing'*, thus, is not determined by a single underlying factor as posited by ATOM model.

In an attempt to trace the development of auxiliary syntax of English speaking children, Theakston and Rowland (2009) conducted a longitudinal elicitation study involving twelve children from the age of 2;10 to 3;6. The focus of the investigation was on the development of auxiliary *BE* in particular *is* and *are* by analysing the children's response to tasks designed to elicit auxiliary *BE* in declaratives and yes/no and *wh*-questions. In terms of accuracy between *is* and *are*, the analysis revealed higher accuracy of *is* (82.8%) than *are* (54.5%) across the three constructions specified earlier. When analysed within specific construction, the researchers noted differences in the development of both *is* and *are*. The form *is* recorded better performance in interrogatives compared to *are*. Nevertheless, *are* performance was much better in declaratives compared to in interrogatives. In general, however, the use of *is* was far more consistent in all the constructions analysed. The patterns of correct use and errors, suggest that each form of *BE* was acquired separately and children did not possess an abstract relation between different forms of *BE* or between sentence types. The data also did not provide a strong evidence that the children were aware of the relation between forms marked for tense, number and person, thus, did not reflect the existence of innate ability or Universal Grammar (Theakston & Rowland, 2009).

2.3.1.1 Summary of *BE* in First Language Acquisition Research

Several conclusions can be drawn from the studies conducted in the nativist camp, one in particular is in the earlier pattern of inflectional morphemes acquisition; copula and

auxiliary *BE* are often omitted in the obligatory contexts. This is consistent with the much earlier morpheme study by Brown (1973), who in his longitudinal study of 14 English morphemes by three children acquiring English as the first language, reported similar omission; *I happy* (Adam, 2;10), *It messy* (Adam, 2;9) (Brown, 1973).

Becker (2002), based on her investigation on declarative constructions, proposed a semantic-predicate sensitivity on the omission of *BE*. She provided empirical data suggesting that supply/omission rate of *BE* was determined by whether they were complemented by stage-level predicates or individual-level predicates. Copula *BE* omission was discovered to be more prevalent with stage-level predicates in both declaratives and interrogatives, which supports the nativists' claim of the presence of universal linguistic parameters assisting child language acquisition.

Besides the semantic-predicates sensitivity, the finiteness of *BE* was also found to influence the pattern of omissions. Schütze (2004) found that the infinitival *BE* was almost never omitted in comparison to copula and auxiliary *BE*. This according to him was due to the realization of Tense in the specification of verbs. Moscati (2006) also posited the same view with regard to *BE* omission in English negation. According to the researcher omission of *BE* in negatives by children acquiring English was the result of truncation of the Tense feature.

What can be generated from the constructivists' studies of *BE* acquisition is the existence of an order of the acquisition of suppletive morpheme where copula *BE* was more consistently used in comparison to auxiliary *BE* (Theakston & Rowland, 2009), thus, providing evidence of copula *BE* being acquired before auxiliary *BE*. This is consistent with the findings of earlier studies (Pine et al., 2008; Wilson, 2003), which posited the following stages in English verb morpheme development:

Table 2.2: Stages in English Verb Morpheme Development

Stage	Verb form
1	Present Participle (V*ing*)
	Irregular Present of copula *BE* (am, is, are)
2	Progressive Aux/*Be* + V*ing*
	Irregular Preterit of Copula *be* (was, were)
3	Irregular Preterit (V*en*) of lexical verbs
4	Regular Preterit (V*ed*) of lexical verbs
5	Regular Present (V*s*) of lexical verbs
6	Irregular Present (does, has)
	Present Perfect Aux Have + V*en*/*ed*

Brown (1973) and Dulay et al., (1982)

Theakston and Rowland (2009) went even further by focusing their investigation on the differences in the acquisition of auxiliary *BE* in particular the pattern of accurate supply and errors of *is* and *are*. The findings revealed a consistent pattern of *is* being more accurately used and recorded less errors compared to *are*. This led the researchers to conclude that such a pattern was the result of higher frequency of *is* in the input, thus, supported the constructivists' argument that elements such as frequency in the input influences early child language acquisition development. The asymmetrical use of *is* and *are* also suggested that there was no innate abstraction at work in the child language acquisition, as the variable supply of the same inflectional morpheme (auxiliary *BE*) provided evidence that *is* and *are* were acquired separately and that the children were not aware of the relation of tense, number and person.

Another finding supporting the constructivists' stand was the subject sensitivity in the provision of *BE*. Wilson (2003) and Pine et al. (2008) recorded significant production of copula and auxiliary *BE* with pronominal subjects as compared to with lexical NP, supplying proof that tense-marking morphemes were acquired in chunks or were very dependent on lexically specific constructions such as *'It's+NP'* and *'He's V-ing'*; the constructions most prevalent in the input. The researchers stressed that these findings provide strong evidence that the verbal morphemes were acquired individually, thus, separately.

2.3.2 *BE* in Second Language Acquisition Research

In the domain of Second Language Acquisition (SLA), the interest on *BE* has always been a major part of investigation of child learner acquisition of the English inflectional projection (IP). Lakshmanan (1995) in her investigation of a longitudinal language data of Marta, a Puerto Rican girl, who moved to US at the age of 4;5, found that after two and a half months of exposure to English the child began to produce copula *BE* in the very early samples (*Mother **is** Mary Jo Fuster* (S1)/*This **is** big bird* (S2)) (p. 58). More interestingly Lakshmanan (1995) noted that in Marta's early production of copula *BE* was never omitted. Auxiliary *BE* also occurred very early in the data and the use bore the existence of functional structures that was inherent in the child's grammar. Through elicitation imitation task Marta was observed to convert contracted auxiliary into uncontracted form. However, these findings did not lend a strong support to the existence of IP system in Marta's production. Although the child had consistent production of both copula and auxiliary *BE*, she was lagging in the production of 3sg present tense *–s* in obligatory contexts.

Child L2 acquisition research has also provided evidence of copula *BE* being acquired earlier than auxiliary *BE* in child IP system. Haznedar (2001) in her longitudinal investigation on the development of English IP system of a 4-year-old Turkish child Erdem, found copula *BE* was among the first verb to appear. Auxiliary *BE* appeared at almost the same time as copula *BE*, however, the supply was more sporadic and the researcher also recorded more omissions of auxiliary *BE* (e.g. *Newcastle going* [S 5]). The researcher concluded, although both copula and auxiliary *BE* occurred at almost the same time, copula *BE* was used "more predominantly" between the two (Haznedar, 2001; Haznedar, 2007). The findings from the emergence of early use of copula and auxiliary *BE* supported the claim of the existence of functional categories in child L2 early acquisition stage and the existence of IP system during that stage.

The acquisition of *BE* was also investigated in the light of finiteness such as the study conducted by Ionin and Wexler (2001). The study involved 20 L1-Russian children aged between 3;9 to 13;10 with one to three years of exposure to English language prior to the study. The findings revealed a disassociation of the supply of copula *BE*, auxiliary *BE*, 3sg present tense *–s* and past tense *-ed*. The researchers noted the existence of a continuum of copula *BE* being more frequently supplied followed by auxiliary *BE* and the least supplied was 3sg *–s*. The pattern was observed in the percentage of morpheme omissions recorded.

Their subjects were also found to produce overgeneration construction *BE + bare V*, in place of progressive participles such as *…the lion **is go** down/and then the police **is come** here* (p. 110). After a detailed examination of all the instances of overgeneration utterances, the researchers concluded that majority of the utterances were not intended as progressives, they instead had generative and stative meaning, as well as past and future meaning. This explained the absence of *–ing* form from the main verb, demonstrating that the learners did not misuse *–ing*, instead *BE* was inserted as a mechanism to mark tense and/or agreement on the main verb. The researchers emphasised that these constructions were clear attribution of poor morphological mapping, whereby learners were unable to access the appropriate affixal inflection and resorted to the use of defaults (uninflected verbal form or suppletive inflection).

Similar to the findings of Ionin and Wexler (2001), Fleta (2003) in her longitudinal investigation of *is*-insertion among four L1 Spanish children acquiring English also found *BE + bare V* constructions in the learners' early acquisition data. The researcher argued that the *is*-insertion construction was a language learning strategy employed by the learners to overcome the difficulty they experienced with raising of verbs, adding that the construction was an economical solution to the intricate process of verb movement operation. *Is*-insertion was also found to be systematically utilised to mark

past tense events (*Andres **is** no want to sleep in the bus*), to express generic/habitual meaning (*The paper **is** not put in the bin*), and was attached to stative verbs which do not require *–ing* form (*The boy **is** no have it*) (p. 89).

The researcher, concluded that *is*-insertion in declarative, negative and interrogative sentences found in her data suggested that learners were constrained by "principles of learning specific to grammar..." (Fleta, 2003, p. 94). The data pointed to a systematic increase in the acquisition of movement operation in English, initially with the insertion of *is* in marking the syntactical and morphological information then moving gradually to the insertion of inflectional morphemes in declaratives and the construction of Subject-Aux Inversion in interrogatives.

Interest in the acquisition of *BE* as L2 language has also been placed on the predicate semantics of *BE*, investigating the influence of individual/stage-level predicates on the acquisition of *BE*. Gavruseva and Meisterheim (2003) conducted a longitudinal study of five children learning English as L2, they obtained almost similar results in the frequency of *BE* omissions with the findings obtained by Becker (2002). Nevertheless, Gavruseva and Meisterheim (2003) argued against the stage/individual-level predicates sensitivity put forward by Becker (2002). Their findings revealed, even though L2 learners omitted *BE* more often before stage-level predicates, the difference in the instances of omission of *BE* before individual-level predicates was too insignificant to prove that individual/stage-level predicates were affecting the learners' supply of *BE*. The overall percentage of null *BE* for individual and stage-level predicates stood at 5% and 9% respectively, which differed significantly with the result obtained by Becker (2002); 56% and 78% for null *BE* in the category of adjectival and locative predicates respectively. Gavruseva and Meisterheim (2003) accounted the differences to the possibility that Becker's (2002) results might be fluctuated due to the miscategorising of the *BE* omission utterances.

Following the nativists' stance on the nature of L2 acquisition, Hawkins and Casillas (2008) offered an explanation on the disparity in the supply of English inflection, namely copula *BE*, auxiliary *BE*, affixal regular past *–ed* and 3sg present tense *–s*, among L2 learners. They argued that the mental grammar of early L2 learners was structured in the same way as the native speakers. L2 learner's inability to produce appropriate inflection according to the nativists was due to the slight difference in the phonological exponents in the Vocabulary entries for verb morphology. While native speakers' entries on "Vocabulary items are specified in terms of bundle of feature of the point of insertion with limited context-sensitivity" (p. 608), L2 learners' entries for exponents are determined by the terminal nodes with which an exponent co-occurred and the insertion is context-sensitive (Hawkins & Casillas, 2008).

In order to test the hypothesis the researchers conducted a completion-task experiment on 20 lower intermediate proficiency speakers of English; 10 L1-Spanish and 10 L1-Chinese. They found copula *BE* /(ɪ)z/ was supplied more than /s/ with simple subject and there was no overgeneralisation of both /(ɪ)z/ and /s/ when the subject was plural. They also found that /(ɪ)z/ and /s/ were consistently supplied even with complex subject with a preceding genitive DP. The findings provided support to the Contextual Complexity Hypothesis, which could be traced through the disassociation of the supply of copula *BE* and auxiliary *BE*. The researchers explained that this disassociation was the result of the constraint learners experience in inserting the Vocabulary item.

Acquisition of *BE* in SLA has also been investigated within the generative framework (White, 1989) in which focus was invested on determining the availability of Universal Grammar (UG) in L2 learners' language faculty. Muneera and Wong (2011) tested the Missing Surface Inflection Hypothesis (Prevost & White, 2000) by analysing adult Arab ESL learners' acquisition of the English functional categories of non-past tense and agreement. The morphemes investigated included third person singular *–s*, and *BE*

auxiliary forms, *is*, *am* and *are*. The participants were 77 undergraduate students from two universities in Yemen. Data were collected using an oral production task (ORPT), whereby learners were asked to orally narrate a story based on a given stimulus. The findings revealed that Arab ESL learners had only managed to obtain approximately 29% correct use non-past auxiliary forms compared to about 55% correct use of non-past thematic verb. Omission of auxiliary *BE* was recorded at approximately 27% and wrongly inflected (WI) items in obligatory contexts was recorded at 43%. The errors identified for WI included, wrong *BE* auxiliary form (inappropriate number), *-ing* deletion, wrong tense, substitution (past tense thematic verb in place of non-past *BE* auxiliary and non-finite of thematic verb in place of non-past *BE* auxiliary) and overgeneration of *BE* forms. The researchers attributed the errors to negative interlingual transfer, explaining that the syntactic difference in English and Arabic contributed to learners' confusion.

Investigation on the development of *BE* among L2 learners, was also conducted in light of interlingual transfer. Lee and Huang (2004) conducted a study on the acquisition development of *BE* among 270 Hong Kong primary school ESL learners aged between 9 to 10. Their main interests were to discover firstly, the system that was inherent in the Chinese ESL learners' interlanguage (IL), secondly the variability of the IL and finally the extent learners' L1 (Chinese) influenced the use of English *BE*. The ESL school children were assigned a story-writing task with a given opening sentence, which they had to complete within an hour. In order to obtain a balanced profile of learners' use of *BE*, the researchers analysed the correct as well as the incorrect use of *BE*.

The findings revealed a developmental pattern similar to the pattern reported in Haznedar (2001); that copula *BE* was acquired earlier than auxiliary *BE*. The ESL learners exhibited confident and stable use of copula *BE* with 80% correct use, compared to only 10% correct use of auxiliary *BE*. Copula *BE* was recorded to occur

in four constructions; *BE-adjective* (e.g. *the king was angry*), *BE-noun* (e.g. *I am a king*), *BE-preposition* (e.g. *the queen is in the palace*) and *BE* in *wh*-question (e.g. *Who are you?*), with 92% correct use of *BE–noun* compared to the other copula *BE* constructions (Lee & Huang, 2004). The finding suggested existence of finer developmental patterns in the acquisition of copula construction with *BE-noun* structure being more easily acquired compared to the other three structures.

The researchers also suggested that the pattern of the correct use of *BE* could be the result of the learners' L1 (Chinese) transfer. They explained that Chinese has a copula-like verb *shi* that links a subject to a noun or noun phrase, but the verb could never be used to link a subject to prepositional and adjective phrase. This explanation justifies the 92% correct use of *BE-noun* structure compared to 61% correct use of *BE-preposition* structure, but could not justify the correct use of *BE-adjective* construction, which was recorded at 74%, when Chinese *shi* "cannot be used to link any predicative adjectives or prepositional phrase..." (Lee & Huang, 2004, p. 213).

With respect to the pattern of incorrect or inappropriate use of *BE*, the findings from the study revealed the same types of errors already attested in the literature, namely omission, overgeneralisation, substitution, subject-verb disagreement and wrong word order. Analysis showed that linguistic context influenced the omission of *BE*, where it was found that learners had the tendency to omit *BE* in a structure *be + very/not/so + adjective* when the adjective was modified by degree adverbs or negated by *not*. The learners were also observed to overgeneralise copula *BE* with a main verb as in *The queen is walked into her bedroom* (Lee & Huang, 2004, p. 218).

Similar to Lee and Huang (2004), Chan (2004) had also focused her investigation on interlingual transfer. The researcher investigated the syntactic transfer of Chinese grammar in the interlanguage of 710 Hong Kong Chinese ESL learners. The data were

obtained from individual interviews, translation and grammatical judgment tasks. 5 error types; incorrect use of copula *BE*, adverbs, inability to use existential *there BE*, relative clause and confusion in verb transitivity were analysed. Chan (2004) found learners tended to omit copula *BE* in the position after modal auxiliaries in the translation task given (46%-47% lower-intermediate group and 18%-11% upper-intermediate group) and they were also unable to identify sentences where *BE* was omitted as ungrammatical (74% lower-intermediate group and 39% upper-intermediate group). Chan (2004) attributed the learners' inability to perform well in the translation and grammatical judgment task to negative interlingual transfer from Chinese grammar explaining that the erroneous structures bore resemblance to normative Chinese grammar. Negative interlingual transfer can also be traced in the learners' performance in the other three error categories; incorrect use of adverbs, existential *there* and relative clause. In the individual interviews learners reported the tendency to think in Chinese; between 73% to 75% of the learners admitted to using Chinese as the thinking language and this according to the researcher confirmed the influence of Chinese in the learners' English.

Unlu and Hatipoglu (2012), have also analysed acquisition of copula *BE* in relation to learners' L1. The subjects of their study were native speakers of Russian who were randomly chosen groups of students from state Russian schools in Moscow. The main objective of the study was to find out if Russian learners faced difficulties while learning English copula *BE* in present simple tense (PST). The researchers explained that copula-type verb as a rule is omitted in PST (the verb has no present tense) and Russians would be producing constructions such as *I-sales clerk*, *My mother-teacher*, *He-interesting* to express PST (Unlu & Hatipoglu, 2012, p. 258). The syntactic difference according to the researchers might negatively influence learners' use of English copula *BE*. The results from two diagnostic tests; multiple choice and item

completion, however, did not suggest negative interlingual transfer as the main cause of learners' difficulties. Even though the Russian learners encountered difficulties learning and applying the correct use of copula *BE*, the types of errors found were developmental. The errors produced most frequently were misformation type errors, whereby learners would either misuse the form of copula *BE* (*You **is** a good tennis player)* or replace *BE* with another verb most often auxiliary verb *do/does* (*The children **do not** at home now*) (Unlu & Hatipoglu, 2012, p. 265). Misformation errors argued the researchers, suggested incomplete understanding and application of the rule of copula *BE* rather than negative interlingual transfer. Furthermore, there were not many omission errors recorded suggesting that learners did not resort to direct transfer of Russian grammar to English. The researchers concluded that despite the syntactic difference between Russian and English with regard to copula *BE* in PST, L1 transfer was not a major cause of errors in the Russian learner data.

In contrast, Murad and Khalil (2015), in their investigation on the errors in English writing committed by L1-Arabic learners, which include omission of auxiliary and copula *BE*, attributed learner errors to interlingual transfer. The learners were found to drop auxiliary *BE* before *Ving* as in *"They ∅ **writing** a story*" instead of *"They **are writing** a story*" and omit copula *BE* to produce construction such as "*he ∅ a strong man*" instead of "*he **is** a strong man*" (p. 478). The researchers argued that the variability in the use of *BE* was the effect of negative transfer from Arabic. According to the researchers copula-like verb is not available in Arabic and the tendency to rely on Arabic when writing had resulted in the transfer of structures from Arabic to English.

SLA researchers, studying interlanguage (IL) grammar also observed that L2 learners often overused *BE* with unaccusative verbs in a structure similar to passive voice as in *BE + Ven.* This construction was well documented in the Chinese interlanguage by Yip (1994), who noted that the unaccusative verbs that were normally passivised belong to a

particular semantic class usually defined by change of state or location and lack of volitional control such as "*What **was happened** yesterday/The leaves **were fallen** down*." (p. 136). Yip (1994) ruled out L1 transfer as the cause of the construction, arguing that such structure would not have been permitted in Chinese grammar. The problem seemed to be universal as the same construction was also produced by learners from other L1 backgrounds (e.g. Arabic, Thai, Japanese, Korean, Italian, and Spanish). The difficulty or confusion learners had with unaccusative verbs, according to Yip (1994) might be contributed by cognitive factor. The researcher argued that learners might have a misguided intuition that there was perhaps a missing agent in the construction with unaccusative such as in *The ship sank* that was interpreted as the ship sinking itself. Yip (1994) explained that learners were inclined to supply the causal agent to make sense of the construction.

Oshita (2000) in her investigation found the same overuse of *BE* similar to the ones documented in Yip (1994), which the researcher termed as overpassivisation due to its similarity to English passive. Through corpus-based investigation of 3362 essays of Italian, Spanish, Japanese and Korean learners of English from Longman Learners Corpus, the researcher found that the most common passivisation errors occur with unaccusative verbs. They were realised in *BE + Ven* structure such as in "…*they **were happened** a few days ago*" (Oshita, 2000). The researcher accounted this instance for the overt evidence of learners' attempts to mark NP movement. In English, passive is often marked by the movement of NP and the use of *BE + Ven* structure to do so. Oshita (2000) argued that L2 learners of English overgeneralised this rule and extended it the unaccusative verbs in their interlanguage.

Ju (2000) conducted a similar study on overpassivisation errors, however, involving a smaller sample of participants; 35 L1-Chinese, all of whom were categorised as advanced learners of English. The researcher found similar results as Yip (1994) and

Oshita (2000) with regard to the class of post-*BE* verbs learners were likely to overpassivise. In the study, unaccusative verbs were found to be likely passivised with the overgeneration of *BE + Ven* structure. Rather than applying syntactic analysis to the errors, Ju (2000) took into consideration the cognitive factors involved in the construction of such errors. The researcher hypothesised that the choice to passivise unaccusative depended mostly on whether or not the learners were able to conceptualise the agents in the discourse. When an agent or cause is/may be a part of learner's mental representation (externally caused events) the unaccusative verb would be passivised and may not when the agent is not clear (internally caused events). The results seemed to indicate that the choice learners made to overpassivise relied heavily on the conceptualisation of the agency, thus, the overgeneration of *BE* with unaccusative verbs was not the result of impaired grammatical knowledge, but rather a result of misapplication of passivisation rules.

Interest in the acquisition of *BE* has also been vested in the effect that formal instruction has on the acquisition of the inflection. Tode (2003) investigated the effects of implicit teaching on the supply of copula *BE* among Japanese learners of English. The participants were 111 Grade 8 and 9 Japanese students, who prior to the study had a year to two years of exposure to English language in the classroom settings. Through written elicitation test, the researcher obtained the data for the supply of *BE*. Analysis of the data uncovered that more than half of the students from both grades failed to supply copula *BE* in obligatory contexts. Similar to the other studies in the literature, she found learners producing sentences such as "*My father a teacher"* whereby *BE* was omitted and also sentences containing overgeneration of copula *BE* with bare *V* as in "*He is like music"* (Tode, 2007, p. 15).

In determining whether the supply of *BE* was the result of analysed versus unanalysed chunks, learners' supply of *BE* was also investigated against other linguistic contexts,

namely the contrast between pronoun subject and noun phrase subject, singular subject versus plural subject, and declarative versus interrogative constructions. In the pronoun/noun phrase hierarchy, the learners were observed to supply higher percentage of correct *pronoun + BE* sequence. According to Tode (2003), it was due to the higher input of *pronoun + BE* learners received, therefore, was easier to memorise as chunks. Another explanation rendered was that unlike nouns the plural formation of pronouns did not involve any suffixation. In the case of nouns determining the correct morphological form of *BE* for them was not as straight forward, as the learners would have to first determine if the noun was singular or plural before they can apply the agreement rule. In the context of singular/plural hierarchy and declarative/interrogative hierarchy correct supply of *BE* for singular and plural pronouns was much higher than singular and plural noun phrase and declarative context recorded higher correct supply of *BE* compared to interrogative context. Tode (2003) explained that the supply of *BE* for singular nouns in declarative involved a straight forward application of the supply rules, while for plural nouns and interrogative required not only application of supply rule, but also agreement and inversion rules (p. 55), which made it harder for the learners to process.

Studies on features of varieties of English for instance African American Vernacular English (AAVE), Creole English and New Englishes have also included investigation on the variability of *BE*. Herat (2005) in his study of Sri Lankan variety of English, discussed zero copula *BE* and compared the variation found in Sri Lankan English with other varieties of Englishes for instance Singaporean English, Malaysian English and AAVE. The study also intended to determine the factors influencing zero copula in Sri Lankan English. The data were gathered from interviews with 18 habitual speakers of Sri Lankan English who were at different levels of proficiency. Using the grammatical environment already attested in previous studies, the absence of *BE* in the Sri Lankan

English data was examined in relation to the type of complement, type of subject and preceding phonological environment.

Zero copula for individual speakers was found to favour *are* absence (17.2%) in comparison to *is* absence (2.4%). The environment in which *are* was mostly absent was in the future environment after *going to* (54%) followed by before adjective phrase (17%), past participle (16%) and *V +ing* (15%). There was no *are* absence recorded in locative position and very low in noun phrase environment (1%). In terms of the preceding phonological environment, zero *are* was found to favour preceding vowels (10.7%) in comparison to the consonants (6.5%), while zero *is* appeared to favour consonants (1.6%) than vowels (0.7%). Zero copula with subject environment in the Sri Lankan English appeared to favour nouns compared to pronouns. Personal pronouns *you, we, they* recorded higher *BE* absence than other pronouns, namely *this, those* and dummy subjects *it, that* and *what*. Based on these findings the researcher concluded that type of complements in particular future marker *going to* and adjective predicate, preceding vowel and NP subjects had significant effects on *BE* absence in Sri Lankan English.

Similar to Herat (2005), Akande (2013) has also discussed the variability in the supply of *BE* in the features of a variety of English specifically the English of Nigerian University Graduates (NGUs). In his findings, Akande (2013) reported two most common non-standard syntactic features of the NGUs' English concerning *BE*; *BE* deletion and intrusion. Deletion of *BE* according to the researcher is the more prominent feature in the NGUs' English compared to intrusion of *BE*. Deletion of *BE* recorded the highest number of instances (113/421). Further analysis of the deletion instances in context revealed that both copula *BE* (***They Ø also in the same category*** *of question but here you have different grammar*) and auxiliary *BE* (*Facilities are all neglected, they are not uh-* ***they Ø either broken down*** *or are obscured, you know.*) (p.

22) tended to be deleted by the participants.

As for *BE* intrusion, the researcher noted that it most often took *BE + base V* sequence (e.g. *are talk, is wish, is sell*), which could also be impaired progressive. Compared to the instances of *BE* deletion, which the researcher reported as being evident in the spoken productions of all the 30 respondents, intrusion was less common. The respondents' high proficiency in English, according to the researcher accounted for the lesser occurrences of *BE* intrusion in NGUs' English. Although the study has managed to present some important features of the non-standard syntactic features in the NGUs' English, it unfortunately did not discuss the possible factors that could influence these non-standard syntactic features.

2.3.2.1 Summary of *BE* in Second Language Research

Several conclusions can be drawn from the studies of *BE* within the boundary of SLA. Firstly, there exists a continuum of order in the acquisition of inflectional morphology in the child L2 acquisition. Copula *BE* is found to be acquired before auxiliary *BE* similar to the order attested in L1 child acquisition (Brown, 1973; Dulay et al., 1982). Child L2 learners in the studies conducted by Lakshmanan (1995), Haznedar (2001), Ionin and Wexler (2001) were discovered to produce copula *BE* earlier than auxiliary *BE*. Even when auxiliary *BE* appeared almost at the same time as copula *BE*, the latter was more consistently supplied compared to the former (Haznedar, 2001; Ionin & Wexler, 2001). In addition, there was also empirical evidence of more confident and stable use of copula *BE* than auxiliary *BE* among older learners of English (Lee & Huang, 2004). These findings lend a strong support to the universality of L2 language acquisition, whereby the child L2 learners in the studies reviewed (Lakshmanan, 1995; Haznedar, 2001; Ionin & Wexler, 2001) regardless of the native languages displayed almost similar path in the acquisition of English functional categories.

Secondly, L2 learners production of *BE* can be influenced by several factors which include the syntactic environments preceding and proceeding *BE* (Gavruseva & Maisterheim, 2003; Hawkins & Casillas, 2008; Herat, 2005, Ju, 2000; Oshita, 2000; Tode, 2003; Yip, 1994), L1 transfer (Akande, 2013; Chan, 2004; Lee & Huang, 2004; Muneera & Wong, 2011; Murad & Khalil, 2015; Unlu & Hatipoglu, 2012), and explicit teaching (Tode, 2003).

The syntactic environments analysed and discussed in the literature include firstly the type of predicates complementing *BE*. The supply of *BE* according to the findings of previous studies could be sensitive to the proceeding predicatives. Gavruseva and Maisterheim (2003) reported higher occurrences of *BE* omission before stage-level predicates than before individual-level predicates. Herat (2005) found that Sri Lankan English speakers tended to omit copula *BE* preceding adjective predicates, while Chinese learners were found to be more confident with *BE-noun* construction compared to *BE-preposition* and *BE-adjective* constructions (Lee & Huang, 2004).

Besides predicate sensitivity, overt or covert *BE* was also found to be influenced by the type of subject preceding it. Tode (2003) reported that Japanese learners tended to supply higher correct *pronoun + BE* sequence compared to *noun + BE* sequence. Similarly Sri Lankan English speakers were also found to omit *BE* more frequently in the *noun + BE* sequence (Herat, 2005). Tode (2003) argued that higher *pronoun + BE* sequence in the input learners received and the complexity of applying supply and agreement rule to noun subjects might contribute to this finding.

The variability in the supply of *BE* was also associated to a special class of intransitive verb that is unaccusative verb. Yip (1994), Oshita (2000) and Ju (2000) found that learners had the tendency to insert *BE* before unaccusative verbs and produced a passive-like construction (*BE + Ven*). Yip (1994) and Ju (2000) explained that the

tendency could be due to cognitive factor, whereby learners could have assumed that an agent was missing in the unaccusative verb construction, thus, provide causal agents to the agentless verb. The overpassivisation according to Oshita (2000) could also be the result of overgeneralisation of passive rule, which was extended to unaccusative verb.

Another influencing factor to L2 learners' production of *BE* is interlingual transfer (Odlin, 2003). The availability of copula-like verb in learners' L1s is believed to influence learners' ability to correctly use *BE*. Lee and Huang (2004) claimed that Chinese learners' ability to correctly produce *BE + noun* construction was a direct influence from Chinese syntax *shi + noun,* however, *shi* could not be used to link a subject to an adjective or a prepositional predicate, which according to the researchers explained the learners inability to correctly produce *BE + adjective* and *BE + preposition.* The tendency learners to think in their L1 according to Chan (2004) also contributed to the transfer. Chinese learners in Chan (2004) admitted to using this strategy and the erroneous structures they produced had traces of normative Chinese grammar. Muneera and Wong (2011) and Murad and Khalil (2015) also attributed Arabic learners' errors in the use auxiliary *BE* in expressing continuous actions to negative interlingual transfer. They argued that in Arabic continuous action is expressed using the non-past form of thematic verb, thus, learners had the tendency to apply the same rule when producing continuous actions in English resulting inappropriate use of auxiliary *BE* (Muneera & Wong, 2011). Unlu and Hatipoglu (2012), however, rejected the L1 transfer factor, arguing that the Russian learners, whose L1 does not have copula-type verb produced more developmental errors. If L1 transfer was the cause of errors, Russian learners should have committed more omission errors, instead they produced misformation errors that suggested incomplete understanding and application of English structure (Unlu & Hatipoglu, 2012).

Moreover, the similar types of errors produced by ESL learners of different L1 backgrounds suggest that many of the errors are developmental. Sri Lankan speakers of English for instance produced similar *BE* omission patterns as the ones found in Malaysian and Singaporean English (Herat, 2005). Insertion of *BE* was also found to exist in many L2 learners' language data including Russian (Ionin & Wexler, 2001), Chinese (Ju, 2000; Lee & Huang, 2004; Yip, 1994), Italian, Spanish, Japanese and Korean (Oshita, 2000) proving that they could have been the outcome of developmental aspect of acquisition rather than L1 transfer.

The literature has also attested several major variability in the use of *BE* among L2, they include two major types of error, namely insertion and omission of *BE*. Ionin and Wexler (2001), Fleta (2003) and Tode (2003) found learners produced *BE + bare V*, whereby *BE* was inserted before a base form of a main verb. According to Ionin and Wexler (2001) *BE* was used as a mechanism to mark tense and/or agreement on the verb. However, Fleta (2003) argued that insertion of *BE* was the result of an economical solution learners took to overcome the difficulties they experienced with verb movement operation. Yip (1994), Oshita (2000) and Ju (2000) found another type of insertion error, where *BE* was inserted before a past participle form of unaccusative verb resembling a passive (*BE + Ven*). Oshita (2000) explained that the construction was an overt evidence of attempts learners took to mark NP movement similar to movement involved in constructing passive. Yip (1994) and Ju (2000), however, related this to the cognitive factor, explaining that such construction was the result of learners assigning an agent to unaccusative verb which does not require one.

Omission of *BE* based on the literature reviewed occurred at almost every stage of the acquisition process. It was a very common feature of child L2 language (Ionin & Wexler, 2001; Haznedar, 2001; Gavruseva & Meisterheim, 2003; Tode, 2003) and a usual occurrence in the language of older ESL learners (Akande, 2013; Herat, 2005; Lee

& Huang, 2004; Muneera & Wong, 2011; Murad & Khalil, 2015). Omission of *BE* among young children was often associated with the acquisition of grammatical morpheme (Haznedar, 2001; Gavruseva & Meisterheim, 2003; Ionin & Wexler, 2001) and implicit teaching (Tode, 2003), while in the case of older learners it was explained by negative interlingual transfer (Lee & Huang, 2004; Muneera & Wong, 2011; Murad & Khalil, 2015) and the influence of the syntactic environments, namely the type of subjects and type of predicates (Herat, 2005).

2.3.3 *BE* in Malaysian English

Platt and Weber (1980) in their description of the features of spoken Malaysian English (ME), found ME speakers to frequently omit *BE* before adjectival (*Kelantan kain sarong Ø very famous.*), nominal (*The house Ø two-storey building.*), locative (*and my brother Ø also in Kedah.*) and verb *–ing* (*Some of them Ø working.*) (p. 74). The supply of copula *BE* was also found to be influenced by the syntactic environment; there was lesser supply of *BE* in pre-locative predicate position and higher degree of supply in pre-nominal predicate position. Platt and Weber (1980) attributed this to negative transfer of the Malay grammar as it has no copula-type verb.

A more recent examination into the grammatical errors in the speech of Malaysian undergraduates conducted by Ting, Mahanita and Chang (2010) found copula *BE* omission and overgeneration were common features in the speech of less proficient learners. Almost 35% of the errors found were omission errors and the majority of them were omission of copula *BE* (*It* ***Ø*** *also good for – for our reading.*). The same pattern was also recorded for overgeneration errors (18%), which were made up mostly of *BE* being inserted inappropriately (*I would like to buy a newspaper, The Sun, umm – which* ***is*** *I heard has an interesting article about Siti Nurhaliza*) (p. 60).

Copula *BE* omission was not only found in the spoken Malaysian English, it was also evident in learners' writing. Maros, Tan and Khazriyati (2007) in their investigation of the factors influencing the acquisition of English among young L1-Malay learners recorded variability in the supply of copula *BE* in the learner essays. In fact copula *BE* omission was found to be one of the major types of errors recorded by the researchers (*My mother's name* ***Ø*** *Maznah binti Hj Dahlan*) (p. 12). Other than omission errors learners were also found to make tense and agreement errors in copula *BE* (*Princess Isabella* ***are*** *very kind and gentle*) (p. 14). These errors were all attributed to negative transfer from the Malay grammar. Similar to Platt and Weber (1980), Maros et al. (2007) argued that copula *BE* is non-existence in the Malay language and *BE* omission was the results of direct transfer from the Malay grammar.

Wee (2009) in her investigation of verb form errors found in L1-Malay ESL learner essays also came to the same conclusion that errors were the result of negative interlingual transfer from Malay grammar. Similar to Maros et al. (2007), one of the major types of errors found in the study by Wee (2009) was omission of *BE*. The omission errors were revealed to occur in the simple present, simple past tense as well as in progressive aspect. Other than omission errors, Wee (2009) also found overgeneration errors, which involved unnecessary addition of *BE* before a main verb as in "*The nurse* ***was bandaged*** *her leg/The accident* ***was happened*** *at Jalan Raja Laut*" (p. 355). The insertion of *BE* before the main verb, in particular the past tense forms *was/were,* according to the researcher could be the result of learners' faulty comprehension of English grammar. They could have interpreted *was/were* as a marker for past tense similar to the function of auxiliary *was/were* in the formation of past progressive (*was/were + Ving*).

Wee (2009) also recorded errors in subject verb agreement involving *BE*. The participants seemed to be confused with English *BE* agreement, opting for plural *are*

with singular subject or vice versa for instance *"The lecturer **are**...", "Plagiarism **are** ..."* and *"Students **is** ..."* (p. 20). The researcher concluded that the complexity of *BE*, with its various inflections and functions may be the contributing factor for learners' confusion. They also did not rule out interlingual transfer as another contributing factor for the variability in the use of *BE* among ESL learners in their study.

Arshad and Hawanum (2010) conducted a corpus-based study investigating specifically the errors in the use of auxiliary *BE* among ESL learners in Malaysia. The study examined the types of errors learners produced when auxiliary *BE* was used and provided suggestions for possible sources of the errors. The findings reveal a common type of errors; the overuse of past forms *BE* in (i) *BE + Ved/en* (*My father **was bought** a dog for me*) and (ii) *BE + bare V* (*My family and I **was go** to Pulau Tioman*…) (p. 169).

The researchers explained that the errors were probably due to the overuse of past tense form *was/were* to indicate past time, while *BE +Ved/en* type errors could also be the result of influence of the positive input that has been incorrectly applied. They gave examples of the use of *BE* before adjective predicate such as *was involved* and the use of *BE + Ven* in passives such as *was caught* and *was kicked.* The learners may have overgeneralised these grammar rules and wrongly applied them. The researchers also added that the *BE + bare V* errors could be the outcome of learners' confusion of the difference between copula and auxiliary *BE*. The learners had treated the *BE* as the main verb and used the past tense form *was* to indicate past time.

The use auxiliary *BE* among Malaysian ESL learners was also investigated in light of the different tenses; present tense (Jishvithaa, Tabitha & Kalajahi, 2013) and past tense (Manokaran, Ramalingam & Adriana, 2013). Both these studies adopted CEA approach in their investigations and both made use of MCSAW (Malaysian Corpus of Students' Argumentative Writing) developed by Mukundan and Kalajahi (2013). Jishvithaa et al.

(2013) found the compositions written by upper secondary school learners contained more correct use auxiliary *BE* in the present tense. However, based on the error patterns concluded that the learners still experienced difficulties with the correct use of auxiliary *BE* in the present tense especial in tense and agreement. As for auxiliary *BE* in the past tense, Manokaran et al. (2013) revealed seven types of errors, namely tense shift, agreement, omission, wrong verb form, addition, misformation and misordering. The findings of these studies suggest auxiliary *BE* in the present is easier for the learners to master than auxiliary *BE* in the past tense.

Based on the findings of the studies reviewed it is clear that ESL learners in Malaysia are facing problems with the correct use of *BE*. They were recorded to produce two major ungrammatical uses of *BE,* namely omission and overgeneration. Omission was found to occur in both copula and auxiliary *BE* constructions and mostly before nominal, adjectival, locative predicates and before present participle *–ing* (Maros et al., 2007; Platt & Weber, 1980; Ting et al., 2010; Wee, 2009; Wee, Sim & Kamaruzam, 2010) and interlingual transfer was believed to be the main influence of *BE* omission. The researchers explained that in the Malay language copula-type verb does not exist and learners' tendency to omit *BE* was the result of direct transfer from the Malay grammar.

Another type of error that was reported across most of the studies reviewed was overgeneration of *BE*. There were two overgeneration patterns recorded (i) *BE + bare V* and (ii) *BE + Ved/en.* Arshad and Hawanum (2010) and Wee, Sim and Kamaruzam (2010) attributed *BE + bare V* overgeneration to past time marking, whereby the past forms of *BE* were inserted to indicate past tense. *BE + Ved/en* overgeneration was believed to be the outcome of overgeneralisation of the English passive rule (Arshad & Hawanum, 2010; Wee, Sim & Kamaruzam, 2010).

2.3.4 Summary of Previous Studies on *BE*

The studies reviewed thus far have given insights into the development of first and second language acquisition research in general and into the acquisition of English IP system specifically. The literature has also presented this development from the perspective of the nativist and constructivist approach. The general consent of both schools of thoughts is that there was an apparent variability in the supply of copula and auxiliary *BE* in the data obtained from first as well as second language acquisition. The nativists contended that the variability in the supply of both copula and auxiliary *BE* showed evidence that the functional categories were present in a child's or learner's grammar. Omission or overgeneration instances may be the result of underspecification of tense and agreement as specified by Agreement and Tense Omission Model (Wexler, Schütze & Rice, 1998) or L2 learners' difficulties with the realisation of the surface morphology (Haznedar & Schwartz, 1997; Ionin & Wexler, 2001, 2002; Lardiere, 1998; Muneera & Wong, 2011; Prevost & White, 2000).

The constructivists (Pine et al., 2008; Wilson, 2003) rejected the unitary category posited by the nativists, they instead argued that verbal morphemes were acquired through the learning of lexically specific constructions. In other words, they maintained that children were not born with the IP system already available at their disposal, but gradually developed the system by learning constructions which happen to contain functional categories (Wilson, 2003). In support of this argument, they presented data in which children's supply of inflectional morphemes was determined by the lexically specific construction such as *he's/that's* (Pine et al., 2008; Wilson, 2003). They also found asymmetrical development in the acquisition of IP system across morphemes and lexical contexts, which provided more ground to support their claim that the acquisition of inflectional system developed in a piecemeal fashion and not bound by an innate abstraction (Pine et al., 2008; Wilson, 2003).

There were crucial differences in the focus of both camps of thoughts, which have influenced the position taken by the present study. The nativists were mostly committed to young children's linguistic development with very limited focus on the relation between linguistic development and other social and cognitive skills like the constructivists. Another fundamental difference was that constructivists not only investigated young children's data, but focused also on older children's and adult's performance providing empirical evidence of the developmental nature of second language acquisition. The findings from the constructivist's camp have most importantly reinforced the idea that language develops through active interaction of the new input learner received with past knowledge and its acquisition would be influenced by social and cognitive factors. This study emulates the SLA theory posited by the constructivists and relied on the principles forwarded by the theory in supporting its findings.

One major concern of researchers of both first and second language acquisition is to determine what influences the variability in the supply of the inflectional morphemes in both L1 and L2 contexts. In the context of L1 research, the supply of both copula and auxiliary *BE* is tied to several factors namely; semantic predicates (Becker, 2002; Moscati 2006), finiteness (Schütze, 2004), subject type (pronominal versus lexical NP) (Pine et al., 2008; Wilson, 2003) and construction type (declaratives/ interrogatives) (Theakston & Rowland, 2009).

Researchers investigating the acquisition of verbal inflection within the domain of L2 acquisition also invested interest in factors such as semantic predicates (Gavruseva & Meisterheim, 2003; Lee & Huang, 2004), finiteness (Schütze, 2004), linguistic contexts (Herat, 2005; Tode, 2003) and post-*BE* verbs (unaccusative) (Ju, 2000; Oshita, 2000; Yip, 1994) in the examination of the variance in the supply of verbal inflections. L2 studies with primary focus on interlanguage grammar and error analysis also discussed

interlingual transfer in their investigation (Akande, 2013; Chan, 2004; Lee & Huang, 2004; Maros et al., 2007; Muneera & Wong, 2011; Murad & Khalil, 2015; Ting et al., 2010; Unlu & Hatipoglu, 2012; Wee, 2009; Wee, Sim & Kamaruzam, 2010). Other factors of interest with regard to the disassociation in the supply of the copula and auxiliary *BE* also include context sensitivity (Hawkins & Casillas, 2008) and explicit teaching (Tode, 2003).

The studies reviewed involved varying degrees of emphasis on *BE*. Studies conducted for instance by Haznedar (2001), Ionin and Wexler (2001), Hawkins and Casillas (2008) positioned *BE* under a larger investigation on the acquisition of verbal inflections, others like Theakston and Rowland (2009), Moscati (2006), Tode (2003) focused on a specific function of *BE* (auxiliary or copula) and then there are researchers like Arshad and Hawanum (2010), Herat (2005), Fleta (2003), whose main interest was only on a particular aspect of *BE* such as the morphosyntactic construction of *is*-insertion (Fleta, 2003). The diverse range in depth and breadth of the investigation on *BE* has resulted in the disparities in the findings especially with regard to factors influencing the patterns of use and misuse.

Even when both the copula and auxiliary *BE* were given equal attention, the methodology undertaken limits the breadth of the investigation, setting invisible boundaries to the aspects of *BE* included in the analysis. Most of the studies reviewed employed elicitation (e.g. Chan, 2004; Hawkins & Casillas, 2008; Ju, 2000; Lee & Huang, 2004) and longitudinal research frameworks (e.g. Fleta, 2003; Gavruseva & Meisterheim, 2003; Haznedar, 2001; Lakshmanan, 1995). They had no doubt contributed significantly to the depth of the investigations, but did not permit for a wider spectrum of investigation to be performed, therefore restricting the research into one or perhaps two focal aspects (e.g. copula/auxiliary vis a vis declaratives/interrogatives) and leaving the other aspects for the next investigation.

Elicitation and longitudinal studies also involved very restricted number of samples or participants making it difficult for the findings to be generalised across other L2 settings.

FLA and SLA studies involved mostly examination on production data of young children acquiring English as their L2 (e.g. Fleta, 2003; Gavruseva & Meisterheim, 2003; Hawkins & Casillas, 2008; Haznedar, 2001; Ionin & Wexler, 2001; Lakshmanan, 1995; Lee & Huang, 2004; Tode, 2003; Unlu & Hatipoglu, 2012). There has been very little focus on the pattern of *BE* in the data of more advanced learners, especially learners who have received considerable amount of formal instruction of English in the context where English is the second language. There is a need to examine their language data and discover if their use of *BE* exhibits the same universality as predicted by the nativists or do they display unique characteristics that might not be accounted by UG?

There is also a need for an investigation involving larger learner samples and more exhaustive investigation on all forms and functions of *BE*, one which is only possible with corpus-based research. Corpus-based study involving large learner corpora will be more exploratory in nature and able to yield properties of learner language that might not be apparent through experimental or elicitation studies. This kind of data is most valuable in the generation of hypotheses about learner language (Barlow, 2005; Leech, 1998).

Finally and perhaps the most important reason to justify the need for this study is the lack of comprehensive investigation carried out that specifically focuses on the use of *BE* by ESL learners in Malaysia. Most of the studies reviewed adopted error analysis approach and focused mainly on identifying and profiling learner errors in which errors in the use of copula and auxiliary *BE* were a part of (e.g. Maros et al., 2007; Siti Hamim & Mohd Mustafa, 2010; Ting et al., 2010; Wee, 2009; Wee, Sim & Kamaruzam, 2010).

To date there have only been three studies specifically investigating the use of *BE* by ESL learners in Malaysia (Arshad & Hawanum, 2010; Jishvithaa et al., 2013; Manokaran et al., 2013) and they were confined to investigating only one function of *BE* (auxiliary) and similar to the error analysis studies conducted earlier (Maros et al., 2007; Siti Hamim & Mohd Mustafa, 2010; Ting et al., 2010; Wee, 2009; Wee, Sim & Kamaruzam, 2010) they were restricted to only analysing learner errors.

Without denying the importance of learner errors as the key to the understanding of learners' interlanguage, investigation on learner language should move beyond error analysis and shift its focus to concurrent analysis of the grammatical and ungrammatical constructions of the target language (TL). There is a need for a comprehensive investigation on the use of *BE* by ESL learners in Malaysia that not only focuses on the errors, but also the correct use of *BE*. The findings from such study could be used to accurately determine what learners have acquired versus what they have not acquired or are in process of acquiring and which aspect of the verb is problematic to them. In addition, findings of the major errors committed by Malaysian ESL learners (Maros et al., 2007; Siti Hamim & Mohd Mustafa, 2010; Ting et al., 2010; Wee, 2009; Wee, Sim & Kamaruzam, 2010) provide evidence of the difficulties the learners experienced with the use of *BE*, thus, to carry out a detail investigation specifically on Malaysian ESL learners' use of *BE* would definitely be a valuable addition to the existing literature.

CHAPTER 3

RESEARCH METHOD

3.0 Introduction

This chapter presents and explains the methodology undertaken to answer the following research questions:

1. What are the similarities and differences in the use of *BE* in the essays compiled in MACLE and LOCNESS?

2. What are the distributional patterns for each form and function of *BE* in the essays written by L1-Malay ESL learners in the Malaysian Corpus of Learner English?

3. What are the patterns of the (a) grammatical and (b) ungrammatical uses of *BE* in the essays written by L1-Malay ESL learners?

4. How do the syntactic environments influence the grammatical and ungrammatical uses of *BE* in the essays written by L1-Malay ESL learners?

The study adopts a corpus-based methodology and involves both quantitative and qualitative analyses of the corpus data. The quantitative analysis involves applying descriptive statistics to the corpus data, while the qualitative analysis takes the form of textual analysis. The quantitative analysis was undertaken to ensure that the main research objectives of the study can be fulfilled. The distributional patterns of the grammatical and ungrammatical uses of all the forms and functions of *BE* in the learner data and the patterns of both the grammatical and ungrammatical uses of the verb can only be obtained by means of quantitative analysis. The qualitative analysis was undertaken mainly to supplement the findings of the quantitative analysis. The study

seeks to find out the patterns of the use of *BE* and the textual analysis provides the insights on how learners use *BE*. The findings from the qualitative analysis provides the explanation, exemplication and interpretation of the use of *BE* in support of the descriptive statistics obtained from the quantitative analysis. Corpus-based investigations according to Biber et al. (1998) "is not only to report quantitative findings, but also to provide functional interpretations of the quantitative patterns" (p. 5), an important principle of corpus-based methodology which the current study aims to emulate. The following sub-sections present in detail the methodology employed and the process undertaken in analysing the corpus data.

3.1 Corpus-based Methodology

The study adopts a corpus-based methodology. It begins with a selection of a set of linguistic categories i.e. the forms and functions of *BE*, prior to the investigation. This according to Biber (2009) sets a corpus-based enquiry apart from corpus-driven investigations. Corpus-based methodology requires access to a digital corpus database and the use of corpus computational tools to annotate and analyse the set of data. This section describes in detail the corpus-based methodology employed for this study. It includes explanation of the following components:

1. Learner Corpora
2. Computational Tools
3. Unit of Analysis
4. Data Coding Procedure
5. Data Analysis Procedure

3.1.1 Learner Corpora

Data are drawn from two learner corpora: Malaysian Corpus of Learner English-MACLE (Knowles & Zuraidah, 2004; Knowles et al., 2006), a Malaysian ESL learner corpus and Louvain Corpus of Native English Essays-LOCNESS (Granger, 1993) an

international native learner corpus. Selection is made based on three key criteria, namely accessibility, representativeness and comparability. It is important to highlight that past studies (see e.g. Leech, 1999; Olofsson, 2004; Tottie & Hoffman, 2006) have highlighted several differences in the grammatical aspects of American and British varieties, which prompted the present study to conduct preliminary analysis on the distribution of *BE* in LOCNESS to determine if there are significant differences in the use of *BE* forms between the American and British English. The results of this analysis is presented in Chapter 4.

3.1.1.1 Malaysian Corpus of Learner English (MACLE)

Malaysian Corpus of Learner English (MACLE) was developed under the supervision of a team of researchers; Gerry Knowles, Zuraidah Mohd Don, Jariah Mohd Jan, Rajeswary Sargunan, Janet Yong, Sathia Devi, Asha Doshi and Su'ad Awab (Knowles et al., 2006) from the University of Malaya, Kuala Lumpur. The corpus consists of approximately 800,000 words accumulated from argumentative essays written by second to fourth year undergraduates from the same university. The essays were collected between 2004 and 2005.

MACLE was developed to represent the Malaysian learner English in the International Corpus of Learner English (ICLE) (Granger, 1993). It was, therefore, designed following the criteria set for ICLE. The requirement for each contribution to ICLE is specified at 200,000 words target. The target should be obtained from a minimum of 200 students since each student is only allowed to contribute a maximum of 1000-word essay each. MACLE has surpassed the target requirement set for ICLE and at the time this study was conducted, the corpus stood at approximately 800,000 words.

ICLE strongly recommends that collected essays be accompanied with a learner profile, which contains information on the learners' demography, their mother tongues and

exposure to English. MACLE has collected the learner profiles as suggested, as well as additional information such as learners' performance in major English examinations, their exposure to English prior to university studies and their socio-economic backgrounds.

Essays contributed to ICLE were from two genres, namely argumentative and literary essays. A list of suitable titles for the essays is available at the Centre for English Corpus Linguistics, UCL website, which can be accessed at http://www.uclouvain.be/en-317607.html. The suggested topics include issues on feminism, crime and punishment, money, science and technology, higher education, rehabilitation system, war, patriotism and censorship. The essays can be assigned as an untimed individual work that students can complete in their own time or as a part of an examination where learners will be timed. The use of language reference tools such as dictionaries and grammar books are permitted, but the essays should be entirely their own. Each essay should be at least 500 words long and should not exceed 1000 words. As presented in Table 3.1, MACLE consists of mainly argumentative essays, it has adapted the suggested topics from ICLE and duly fulfilled the other task requirements set by ICLE.

MACLE is selected for this study because it is the only learner corpus in Malaysia that contains a sizeable collection of essays written by ESL undergraduates, who are the target samples of this study. Other learner corpora available in Malaysia, namely the English of Malaysian School Students corpus-EMAS (Arshad, 2002) and Corpus Archive of Learner English in Sabah/Sarawak-CALES (Botley et al., 2005) do not fulfil the research requirement. EMAS corpus is a compilation of spoken and written language of ESL learners from primary and secondary schools, while CALES, although consists of essays written by undergraduates, is comparatively smaller than MACLE. It contains about 400,000 words. Thus, MACLE is deemed more representative of the

Malaysian ESL learners from the specific age and level of English language proficiency set by the study.

Table 3.1: Design Criteria of MACLE

Learner Variables	
Learning context	: University
Learner type	: English as a Second Language
Level of learners	: Non-native English third and fourth year undergraduates
Mother tongue	: Malay, Chinese and Tamil
Mean Age of learners	: 24 years old
Task Variables	
Task type	: Assignment
Text genre	: Argumentative
Essay topics	: Issues on Feminism, Crime and Punishment, Money, Science and Technology, War and Patriotism, Higher Education, Rehabilitation System, and National Security and Censorship
Word limit	: 500 words and more
Use of reference tools	: Allowed
Setting	: Timed and untimed
Years Collected	: 2004 - 2005

In terms of size, MACLE is considered quite small in comparison to other learner corpora such as ICLE that consists of 2 million words (Granger, 1993). Biber (1990) suggests that counts of any linguistic feature would be stable across 1000-word samples from a text (Biber et al., 1998). In the case of MACLE, each text consists of approximately 500 words and more, which is half the ideal sample size suggested by Biber (1990). Nevertheless, the size is considered sufficient for the purpose of investigating *BE*, since it is a very common verb (Biber et al., 1999). *BE* is expected to occur frequently in the learner corpus. In addition, Biber (1990) discovered that many grammatical features were found in only 10 texts from Lancaster-Oslo/Bergen Corpus (LOB) corpus (Biber et al., 1998), suggesting that investigation on common language features such as *BE* can produce stable results even with the use of a small corpus.

In addition, Granger (1998b) adds that learner corpora are not expected "to reach the gigantic sizes of native corpora" (p.10). Firstly, because it is not easy to gather learner data and secondly the optimum size of the corpus depends mostly on the linguistic

investigation to be carried out (de Haan, 1992). In this case, the corpus size is deemed suitable to represent the language of ESL learners in Malaysia, as it surpasses the requirement set by ICLE that only requires a contribution of approximately 200,000 words for each of its national sub-corpus (Granger, 1998b). Table 3.2 below presents the statistical information on MACLE which was generated from WordSmith Tools version 5.

Table 3.2: Statistical Information of MACLE

	MACLE	**L1-Malay Sub-corpus**
Bytes	5,426,823	1,398,040
Tokens	895,204	198,262
Types	20,851	9,951
Type/token ratio (TTR)*	2.35	5.03
Standardised TTR**	38.8	38.49
Average word length	4.58	4.51
Sentences	53,246	11,707
Average mean (in words)	16.66	16.90
Paragraphs	5,760	1,109
1-letter words	24,853	5,697
2-letter words	162,189	36,219
3-letter words	179,550	39,180
4-letter words	164,904	37,252
5-letter words	107,392	22,966
6-letter words	71,608	16,046
7-letter words	62,705	14,221
8-letter words	43,653	9,671
9-letter words	30,792	6,889
10-letter words	26,053	5,390
11-letter words	11,707	2,527
12-letter words	5,022	1,109
13-letter words	2,721	508
14-letter words	1,106	293
15(+)letter words	950	215

*the type/token ratio shows the number of types per 100 tokens
**the standardised TTR shows the number of types per 100 tokens. It is applied to reduce the influence of corpora of different sizes on the ordinary TTR based on 1000 running words of text.

3.1.1.2 L1-Malay Learners' Contribution to MACLE

As mentioned in Chapter 1, the analysis of the use of *BE* in this study will be limited to essays written by L1-Malay ESL learners. Table 3.3 summarises the composition of the L1-Malay learners' contribution to MACLE.

Table 3.3: Composition of L1-Malay Learners' Contribution to MACLE

No	Faculty	Number of Essays
1	Faculty of Built Environment	17
2	Faculty of Business & Accountancy	67
3	Faculty of Computer Science	21
4	Faculty of Economy	28
5	Faculty of Education	53
6	Faculty of Arts & Social Sciences	11
7	Faculty of Islamic Studies	16
8	Faculty of Engineering	54
9	Faculty of Law	44
10	Faculty of Medicine	6
11	Faculty of Science	17
12	Faculty of Languages & Linguistics	31
13	Academy of Malay Studies	1
	Total number of essays	366

There are in total 366 essays contributed by L1-Malay ESL learners from 13 different faculties. The number of essays contributed from each faculty varies; they range from 1 to 68 scripts depending on the size and the number of L1-Malay learners in the faculty. The number of accumulated tokens for the L1-Malay learner sub-corpus is 198 262 words (refer to Table 3.2), which is one fourth of the tokens contained in MACLE.

For the writing assignment, the learners were given the following 10 prompts to choose from:

1. Crime does not pay.
2. The prison system is outdated. No civilized society should punish its criminals: It should rehabilitate them.
3. Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.
4. A man/woman's financial reward should commensurate with their contribution to the society they live in.
5. The role of censorship in society

6. All armies should consist entirely of professional soldiers: there is no value in a system of military service.
7. The Gulf War has shown us that it is still a great thing to fight for one's country
8. Feminists have done more harm to the cause of women than good.
9. In the words of the old song: 'Money is the root of all evil'.
10. Some people say that in our modern world, dominated by science, technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?

Table 3.4 below summarises the learner profile for L1-Malay learner sub-corpus with detailed information regarding the topics learners had written, their age, gender, native language, year of study and years of formal exposure to English.

Table 3.4: Summary of Learner Profile for MACLE

Gender	Female -287 (78.4%) Male – 79 (21.6%)
Native Language	Bahasa Melayu (Malay)
Average Age	24 years old
Average Exposure to English (in school)	11 years
Average Exposure to English (in university)	3 years
Year of Study	1st – 1 (0.27%) 2nd – 81 (22.13%) 3rd – 112 (30.60%) 4th – 154 (42.08%) 5th – 18 (4.92%)
Topic Written	1 – 40 (11%) 2 – 22 (6%) 3 – 88 (24%) 4 – 0 (0%) 5 – 30 (8.2%) 6 – 1 (0.3%) 7 – 6 (1.6%) 8 – 7 (1.9%) 9 – 115 (31.4%) 10 – 57 (15.6%)

A large majority of the learners (82%) had chosen to write topics 1, 3, 9 and 10 and the others (18%) chose topics 2, 5, 7 and 8, only one person chose topic 6, while topic 4 had no takers. The learners' age at the time the essays were collected ranged from 20 to 47

years old with the average age of 24 years old. The older group of learners (30s and 40s) belonged to the Faculty of Education; they were mostly in-service teachers who were offered the opportunity to pursue their first degree in education at the university. They generally had obtained their Diploma in Education and have had years of teaching experience prior to enrolling for first degree at the University of Malaya. This explains the age difference between them and the other learners.

With regard to the learners' gender, there is a significantly higher composition of female learners (287/366) compared to male learners (79/366). As mentioned earlier the majority of the learners were second to fourth year students (94.81%), with an exception of one (1) first year students and eighteen (18) fifth year students. The learners have had an average of 11 years exposure to English in their ESL classrooms in primary (6 years) and secondary schools (5 years). In addition, they also had to enroll in compulsory English language courses at the university for 3 years.

3.1.1.3 Louvain Corpus of Native English Essays (LOCNESS)

Louvain Corpus of Native English Essays (LOCNESS) is a native speaker component of the International Corpus of Learner English (ICLE) (Granger, 1993). It is made up of essays written by British and American university students for their assignments and examinations. Each essay has approximately 500 words and more. The corpus has a total of 324,304 words which is the total combination of the following sub-corpora:

- British pupils' A-Level essays: 60,209 words
- British university students essays: 59,568 words
- American university students' essays: 168,400 words

Only the university students' essays will be included in the analysis of this study, which means that essays from British A-Level pupils will be excluded since these pupils were still at the pre-university level when their essays were compiled.

The university level sub-corpora consist of not only argumentative, but also literary essays. In ensuring comparability between MACLE and LOCNESS, only argumentative essays were selected to represent the native learner language. Table 3.5 presents the actual number of essays selected from LOCNESS.

Table 3.5: Composition of British and American Learner Sub-Corpora

	No of essays	No of words
British Universities*	33	19,019
American Universities:		
1. Marquette University	46	54,285
2. Indiana University at Indianapolis	28	13,454
3. Presbyterian College, South Carolina	6	12447
4. University of South Carolina	53	52885
5. University of Michigan	43	16502
Total	**209**	**168592**

*the names of the universities are not available

Table 3.6 below summarises the statistical information of LOCNESS and its sub-corpora:

Table 3.6: Statistical Information of LOCNESS and Its Sub-corpora

	Ame. & Bri. Sub-Corpora	Ame. & Bri. Argumentative Essay Sub-Corpus
Bytes	1,567,687	1,002,309
Tokens	263,767	168,582
Types	14,753	11,048
Type/token ratio (TTR)*	5.62	6.59
Standardised TTR**	40.13	40.03
Average word length	4.70	4.70
Sentences	13,352	8,822
Average mean (in words)	19.65	19
Paragraphs	671	448
1-letter words	7,032	5,083
2-letter words	48,240	29,44
3-letter words	51,869	32,193
4-letter words	43,420	29,420
5-letter words	28,268	18,512
6-letter words	22,367	14,437
7-letter words	21,377	13,782
8-letter words	15,534	9,884
9-letter words	10,793	6,636
10-letter words	7,608	4,607
11-letter words	3,765	2,527
12-letter words	1,709	1,026
13-letter words	1,094	653
14-letter words	454	247
15(+)-letter words	236	130

*the type/token ratio shows the number of types per 100 tokens

**the standardised TTR shows the number of types per 100 tokens. It is applied to reduce the influence of corpora of different sizes on the ordinary TTR based on 1000 running words of text.

For the purpose of this research LOCNESS serves as the control corpus, representing the native learner English. It mainly functions as a reference corpus in determining the similarities and differences in the patterns of the use of *BE* by the L1-Malay ESL learners in comparison to the NS learners. The comparison, however, does not intend to place the NNS learner language within the use, overuse and underuse framework proposed by Leech (1998) Granger (2002) and Barlow (2005).

3.1.1.4 Comparability of Corpora: LOCNESS and MACLE

LOCNESS is chosen as the native control corpus for this study mainly because it is comparable to MACLE in terms of the learner level of studies and the task variables. Both LOCNESS and MACLE consist mainly of argumentative essays written by university undergraduates. LOCNESS is also a better choice compared to other general native corpora like the British National Corpus (BNC) or Longman Spoken and Written English Corpus (LSWEC), which consist of a very wide range of text types and language beyond what is expected of the ESL learners. Hyland and Milton (1997) and McCrostie (2008) have criticised the use of large native corpus such as BNC or LSWEC to be compared to ESL learner corpora, which according to them would set such a high standard to the ESL learners. ESL learners are not expected to write on wide range of topics and subjects and are definitely not expected to possess the same level of writing standard as the NS writers. It is also not necessary for the learners' writings to be modelled according to the standard of professional writers such as journalists, authors or novelists. It is, therefore, sufficient for the ESL learners' writings to be modelled after the writings of NS learners as recommended by Hyland and Milton (1997) and McCrostie (2008). Table 3.7 presents the comparison of LOCNESS and MACLE.

Table 3.7: Learner and Task Variables of LOCNESS and MACLE

	MACLE	LOCNESS
Learner Variables		
Learning context	University	University
Learner type	ESL	ENL
Level of learners	Non-native English third and fourth Year undergraduates	Native British and American university undergraduates
Mother tongue	Malay, Chinese and Tamil	English
Average age of learners	24 years old	21 years old
Task Variables		
Task type	Assignment	Examination and Assignment
Text genre	Argumentative	Argumentative
Topics	10	100
Word limit	500 words and more	500 words and more
Reference tools	Allowed	Not-allowed
Setting	Timed and untimed	Timed and untimed
Years Collected	2004, 2005	1991, 1995

3.1.2 Computational Tools

This section describes the computational tools utilised in the study. There are 2 main computational tools used; CLAWS4 part-of-speech (POS) tagger to tag the native speaker corpus (LOCNESS) and WordSmith Tools version 5, a text retrieval software program to generate tokens of *BE* in both LOCNESS and L1-Malay learner sub-corpus.

3.1.2.1 Part-Of-Speech (POS) Tagger: CLAWS4

This study makes use of CLAWS4 (Constitute Likelihood Automatic Word-tagging System) developed by UCREL at the University of Lancaster (Garside & Smith, 1997) to tag the native speaker corpus (LOCNESS). It is a hybrid tagger using both the probabilistic and rule-based elements. The probabilistic element in the tagger enables it to select a grammatical or part-of-speech (POS) tag for a word by calculating the likelihood of all the probability of all possible tags to occur in a particular context and choose the tag sequence with the highest probability (Garside & Smith, 1997, p. 103). The frequencies obtained from the calculations are categorised into a three-point scale:

(i) common

(ii) rare (less than 10% of the word occurrence is expected to receive this tag) and

(iii) very rare (less than 1% of the word occurrence is expected to receive this tag)

Garside & Smith (1997, p. 107)

The probabilistic element alone, however, is insufficient to accurately tag a text as it "treats the tag sequence as an abstraction" (Garside & Smith, 1997, p. 105). It is, therefore, unable to accurately assign exceptional coding for sets of words and expressions such as idioms, multiwords or foreign expressions. As a complement to the probabilistic element, CLAWS4 also incorporates a rule-based element. The rule-based element enables idioms such as *as well as*, *in order that* to be tagged as single tokens and compound nouns such as *dining room* as NOUN-NOUN rather than ADJECTIVE-NOUN. The incorporation of both probabilistic and rule-based components in CLAWS has resulted in around 96-97% accuracy in coding the 100 million-word British National Corpus and it is also expected to produce similar accuracy rate across other texts (Garside & Smith, 1997, p. 120).

In view of the high percentage of accuracy, CLAWS is chosen to tag the selected argumentative essays in LOCNESS. By tagging the corpus, it would be easier to extract specific grammatical patterns relevant to this study such as '*BE + intensifier + adjective*', and the extraction can be done automatically. Besides that, more grammatical information could be extracted from the tagged corpora. In addition, POS tagged corpus would be able to reveal features that cannot be automatically detected from the raw corpus (Garside, Leech & McEnery, 1997, p. 4).

Nevertheless, the tagger would not be used to tag MACLE data due the possibility of tagger inaccuracy rooted from learner errors. MACLE, though consists of essays

written by more advanced ESL learners, are unfortunately not error-free. Under this circumstance, CLAWS could not be utilised to tag MACLE, since there are possibilities that the tagger would not be able to accurately tag the errors in the ESL learner corpus. According to Díaz-Negrillo et al. (2010) automatic taggers such as CLAWS are not designed for or trained with L2 learner data, making them unfamiliar with the erroneous structures that are found in learner corpus. Several studies have highlighted the influence of learner errors on automatic tagger accuracy (de Haan, 2000; Díaz-Negrillo et al., 2010; Meunier & De Mönnink, 2001; van Rooy & Schäfer, 2002). According to these studies, errors in spelling (de Haan, 2000; Meunier & De Mönnink, 2001; van Rooy & Schäfer, 2002), lexical choice, verb conjugation, clause type, infinitive, omission (van Rooy & Schäfer, 2002) have very serious influence on tagger accuracy.

van Rooy and Schäfer (2002) in their evaluation of the accuracy of three automatic POS taggers; TOSCA-ICLE, Brill tagger, and CLAWS, in tagging 5 randomly selected essays from the Tswana Learner English Corpus (TLEC) found that two thirds (2/3) of tag errors in CLAWS, a third (1/3) in TOSCA and a quarter (1/4) in Brill, were due to learner errors. In view of the possible tagger inaccuracy, which could be caused by the learner errors, it is decided that MACLE shall be tagged manually.

3.1.2.2 WordSmith Tools

This study employs WordSmith Tools Version 5 (Scott, 2008) to analyse the corpora. This software suite includes three major tools; Concord, WordList, Keywords.

1. **Concord**. It is the most useful tool for the analysis of the syntagmatic relations between lexical items. It can provide researchers with concordances of words, word partials and sequences of words. For example, if one wants to see how many nouns are in a POS-tagged corpus just type "*_NN* (the first asterisk stands for any word occurring before "_NN*, and the second for the scatter of

any forms of NN such as NN1, NN2, NNA. The search item or node will then appear in the center of the concordance lines, allowing for the linguistic patterns around the central node to be observed. The patterns appear in a column block, which enables them to be easily identified. Concord also allows access to other analysis elements, namely collocates, word distance of collocates from the search word, dispersion plot showing where the search item appears in each text and cluster showing the repeated patterns involving the searched word.

2. **WordList**. It can be used to produce word lists or word-cluster lists of a text or corpus. The words or word clusters in these lists can be arranged in alphabetical order or by frequency of occurrences. Moreover, the tool also provides some statistical information such as "token", "type", "type/token ratio", "word length", "percentage of a word" and etc.
3. **KeyWords**. It is based on comparison of the word or word-cluster list. If a word in a corpus occurs more frequently than that in a reference corpus, it is considered as a keyword. The keyword is calculated on the basis of the log likelihood test, which 'gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus (Scott, 1998, p. 72).

For the purpose of this study, only Concord and Wordlist Tools were utilised for data analysis. Nevertheless, the major part of the data analysis made use of mostly Concord Tool, Wordlist Tool was only utilised at the beginning of the analysis to generate the statistical information of the corpora involved in the study i.e. MACLE and LOCNESS.

3.1.3 Unit of Analysis

This section presents the unit of analysis of the study. According to Biber et al. (1998), one of the very important aspects of corpus-based investigation is to determine the unit of analysis of the investigation. It is also termed "observations" and in studies

investigating linguistic structures like the present study, "each observation is an occurrence of the structure in question" (Biber et al., 1998, p. 269).

In setting the unit of analysis of this study, it is necessary to refer to the objectives of the study; (i) to identify the distributional patterns for each form and function of *BE*, (ii) to identify the patterns of the grammatical and ungrammatical uses of *BE* and (iii) to determine the extent of the influence of the syntactic environments on the grammatical and ungrammatical uses of *BE*. In order to fulfill these research objectives, the unit of analysis of the study has to include (i) all the forms of finite and non-finite *BE* in order to obtain the distributional patterns of all the forms and function of *BE*, (ii) all the constituents occurring before (pre-*BE*) and after *BE* (post-*BE*) in order to determine the patterns of the grammatical and ungrammatical uses *BE* and the extent the constituents influence the constructions of the grammatical and ungrammatical uses of *BE*.

The unit of analysis for *BE* in this study comprises a single-verb unit of all finite *BE* (*am, is, are, was* and *were*) and non-finite *BE* (*be, been* and *being)*. The negative forms of finite *BE* namely; *am not, is not, are not, was not* and *were not*, although composed of two linguistic features (*BE + adv*) are considered as single units. As for the pre-*BE* and post-*BE* constituents, the unit of analysis is set based on the analysis parameters presented and discussed in Section 3.1.4.2. The constituents can be single-word units such as "*money", "students", "happy"* or "*evil",* a phrase such as "*the government of Malaysia", "their needs" or "at the university"* or a clause such as "*who studies at a local university", "which nobody recommends"* or *"that he has borrowed".* Table 3.8 presents the division of the unit of analysis for all the linguistic features observed in this study.

Table 3.8: Unit of Analysis of the Study

	Single unit	**Phrase**	**Clause**
Finite *BE*	am, is, are, was, were am not, is not, are not, was not, were not isn't, aren't, wasn't, weren't, 'm, 's, 're	NA	NA
Non-finite *BE*	be, been, being		
Subjects	**Pronouns** **Personal Pronouns** I, you, he, she, it, we, they **Demonstrative pronouns** this, that, these, those **Indefinite pronouns** one, some, all, both, many, no one etc. ***Wh*-subjective pronouns** who, which **Nouns** money, computer, university etc.	**Noun Phrase** A small group of students… The computer courses… Some feminist groups…	NA
Subject predicates	**Pronouns** She is *me* and I am *her*. **Adjectives** He is *smart*. **Adverbs** I was *there* yesterday.	**Noun phrase** The kernel **is** *the part of the plant of greatest value*. **Adjective phrase** That **wasn't** *very nice*. **Prepositional phrase** The houses **are** *in the conservation area.*	**Infinitive *to* clause** The capital **is** *to be provided by the French government*. ***That*-clause** But the danger **was** *that the pound would fall further than planned* ***Wh*-clause** That**'s** *why I bought the refill*
Post- *BE* verb	**Transitive verbs** take, lick, eat, etc. **Unergative verbs** dance, shout, work, cry, laugh, etc. **Unaccusative verbs** melt, happen, sink, appear, break, fall, etc.	NA	NA
Auxiliaries	***have*** has, have, had **Modals** can, may, shall, will, could, might, should, would	**Semi-modal** have to, ought to, be going to, be supposed to, have got to, used to	NA
Intensifiers	**Degree adverbs** very, always, so, extremely, highly, greatly, **Negation** not	NA	NA

3.1.4 Data Coding

3.1.4.1 Analytical Parameters for Forms and Functions of *BE*

Before proceeding with the data coding procedures, it is necessary to clearly set the analytical parameters of the forms and functions of *BE* for this study. The parameters serve as a guide for data analysis as they clearly set the limits of the linguistic items to

be analysed. This enables the analysis to be conducted within the scope of the study. Previous corpus-based studies on grammar for instance Housen (2002) on L2 acquisition of the English verb system adapted Vendler's (1967 cited in Housen, 2002) model of lexical verb aspect and Dulay, Burt & Krashen's (1982) hierarchy of development of verb morphemes in English L2 acquisition as the analytical parameters. A more recent corpus-based study by Zhang (2015), which compares the use of extraposition in academic writing and popular writing, adopted Collins' (1994) four major types of extraposition clause patterns, Herriman's (2000a) matrix for predicate classification and Biber et al.'s (1999) semantic classification of single-word verbs in her analysis. Both these studies highlight the importance for any corpus-based study on grammar to clearly set its analytical parameters as they will determine the analysis boundaries, thus, eliminate the possibility of time being wasted on unnecessary analysis.

In setting the parameters for this study, *Longman Grammar of Spoken and Written English* (LGSWE) (Biber et al., 1999) is used as the main reference, from which the parameters for the forms and functions of *BE* are obtained. LGSWE is chosen as it is the only corpus-based grammar developed from detailed and comprehensive analyses of Longman Spoken and Written English Corpus, a large corpus containing approximately 40 million words. Below are the advantages of a corpus-based grammar, which makes it an ideal reference for the analysis of *BE* in this study:

1. Makes use of only authentic examples available in the corpus,
2. Covers a wide range of language variation; conversation, fiction, writing and news writing,
3. Provides information of speakers' preferences and frequency of use,
4. Includes interpretation of frequency according to context and discourse and
5. Brings together lexico-grammatical relationship.

Biber, Conrad and Leech (2002, p. 3)

A. Analytical parameters for *BE* forms

The analytical parameters for *BE* forms includes all its eight inflections, five of which are finite forms (*am, is, are, was, were*) and three non-finite forms (*be, been, being*). In addition, the analysis also includes the negative and contracted forms of finite *BE*. The list of *BE* forms analysed in this study is summarised in Table 3.9 below:

Table 3.9: Analytical Parameters of *BE* Forms

Finite *BE* Forms	
Base forms	am, is, are, was, were
Negative forms	am not, is not, are not, was not, were not
Contracted forms	isn't, aren't, wasn't, weren't, 'm, 's, 're
Non-Finite *BE* Forms	
Infinitive form	be
Past participle form	been
Progressive form	being

Adapted from Biber et al. (1999)

B. Analytical parameters for the functions of finite *BE*

BE performs multiple functions. One of its primary functions is as a main verb or a copular, which links a subject NP of a sentence to its subject predicate that can be in a form of a phrase (noun phrase or an adjective phrase) or a clause (*that* clause or infinitive *to* clause). It can also be used to link the subject NP to its obligatory adverbial that is usually in the form of a prepositional phrase as exemplified in Table 3.10.

Its other main function is as an auxiliary, which is used to mark progressive aspect by combining *BE* with progressive form of a lexical verb as in "*The children **are playing** in the sand box"* and to form passive voice when *BE* is combined with the past participle of a lexical verb as in "*I **was informed** of the changes in the work schedule by the office secretary"*.

In addition, it also performs the task of a negative operator, which when combined with the negation *not* (*is not, are not, was not* and *were not*) transforms a declarative to a negative. The next function is as an interrogative operator, which is used mainly to

construct, yes/no question in which the *BE* is positioned in the front of the sentence as in "***Are*** *you sick?"* It is also used in question tag, which is normally added at the end of a clause as in "*She is nice,* ***isn't she***?"

Finally, *BE* is also used in the unique constructions of existential *there* structure and *it*-cleft. These functions with their respective examples are presented in Table 3.10 below.

Table 3.10: Analytical Parameters for the Functions of Finite *BE*

Functions		Examples
Copular	Link subjects NP to	
	a. Subject predicate	Radio waves ***are*** useful
	b. Obligatory adverbial	She ***was*** in Olie's room a lot.
Auxiliary	Mark	
	a. Progressive aspect (*BE* + V*ing*)	The last light ***was fading*** by the time he entered the town.
	b. Passive voice (*BE* + V*en*)	This system of intergovernmental transfers ***is called*** fiscal federalism.
Negative operator	*BE* + not	They ***are not*** forgotten. You'***re not*** pretty.
Interrogative operator	Yes/No Questions (*BE* + Subject)	***Is it*** Thursday today?
	Question Tags (*BE* + pronoun subject)	She is so generous, ***isn't she***?
Existential structure – *There BE*	(there + *BE* + noun phrase + place or time position adverbial)	***There are*** around 6,000 accidents in the kitchens in the Northern Ireland homes every year.
It-cleft	It + *BE* + (NP/PrepP/Adv/AdvC)	His eyes were clear and brown and filled with appropriate country shyness. ***It was*** his voice [that held me] ***It was*** only for the carrot [that they put up with his abominable parties]

Adapted from Biber et al. (1999)

C. Analytical parameters for the functions of non-finite *BE*

The non-finite *BE* takes on three forms; *be*, *been* and *being*. These *BE* forms are not considered as main verbs and are generally used in conjunction with auxiliary *have* (*have been*), *modals* (*will be*) and *BE* (*is being*). Nonetheless, when combined with an auxiliary, a modal or a finite *BE* they each performs one or more functions.

The infinitive *be* form, when combined with modal auxiliary as in "*I* ***will be*** *there soon"* or with modal and progressive aspect as in "*She* ***will be attending*** *college next month"* is used to express future state. It is also used in passive voice when combined with a

modal auxiliary and a past participle as in "*Sally **could be charged** under the penal code, if the eyewitness refuses to testify*".

The form *been* is only used after auxiliary *have* and can be used to perform three functions. The first is to form a present/past perfect tense when it is used after auxiliary *have* as in "*I **have been** there before"*. The second is to form perfect progressive when it is combined with auxiliary *have* and a progressive form of a lexical verb as in "*He **has been seeing** a doctor for his cancer treatment"*. Third, when used with auxiliary *have* and the past participle form of a lexical verb, it functions as a perfect passive as in "*He **had been stripped** of all his rights"*.

The form *being* has a very limited function. It is commonly used after a finite *BE* form and combined with a past participle to form progressive passive as in *"The candies **are being sorted** according to colours and sizes".* Table 3.11 summarises the forms, functions and sample sentences of the non-finite *BE* forms.

Table 3.11: Analytical Parameters for the Functions of Non-Finite *BE*

Forms	Functions	Examples
be	future tense (modal + *be*)	Even more precise coordination ***will be*** necessary
	modal in passive voice (modal + *be* + V*en*)	The methods ***could be refined*** and made more accurate.
	modal in progressive aspect (modal + *be* + V*ing*)	Nancy ***will be coming***
been	present/past perfect (have/has /had+ *been*)	Rowlands ***has been*** critical of Welsh officials
	perfect progressive (have/has/had + *been* + V*ing*)	He ***had been keeping*** it in a safety deposit box at the Bank of America
being	perfect passive (had + *been* + V*en*)	He ***had been thrown*** from a moving train
	progressive passive (*BE* + *being* + V*en*)	A police spokesman said nobody else ***was being sought*** in connection with the incident

3.1.4.2 Analytical Parameters for Pre-*BE* and Post-*BE* constituents

In determining if the syntactic environments influence the grammatical use as well ungrammatical use of *BE*, the verb has to be analysed concurrently with its pre-*BE* and post-*BE* constituents. Studies investigating the variability in the use of *BE*, in particular

omission and addition of *BE* among ESL learners of various L1 backgrounds have identified several pre-*BE* and post-*BE* constituents that could influence the variability in the supply of the verb. These constituents include, the types of subjects (Herat, 2005; Tode, 2003, 2007; Wilson, 2003), the types of predicates (Gavruseva & Meisterheim, 2003; Herat, 2005; Lee & Huang, 2004), the types of post-*BE* lexical verbs (Ju, 2000; Oshita, 2000; Yip, 1994) and the presence of intensifiers and modal auxiliaries (Chan, 2004; Lee & Huang, 2004)

The analytical parameters for the pre-*BE* and post-*BE* constituents, which are based on the findings of previous studies, are set with the aid of *Longman Grammar of Spoken and Written English* (Biber et al., 1999). For the types of subjects, the analysis includes any form of lexical nouns or noun phrases and the different types of pronouns that function as the subject NP of a sentence. As for the types of subject predicates they include phrases; noun, adjective and prepositional phrases, and complement clauses; *that*-clause, infinitive *to* clause and *wh*-clause.

The post-*BE* verbs are analysed according to the verb classes proposed by Unaccusative Hypothesis (UH) formulated by Perlmutter (1978). Perlmutter distinguishes intransitive verbs into two finer classes; unaccusatives and unergatives. As a result, the verb class for this study is divided into three; transitive, unergative and unaccusative.

The analysis of pre-*BE* and post-*BE* constituents also includes the presence of auxiliaries; auxiliary *have* and modal auxiliaries. Lastly, the presence of intensifiers is also analysed, the intensifiers under investigation include degree adverbs and negation *not*. Table 3.12 below summarises the parameters set for the analysis of pre-*BE* and post-*BE* constituents with a brief description and examples for each constituent.

Table 3.12: Analytical Parameters for Pre-*BE* and Post-*BE* Constituents

Pre-*BE* and Post- *BE* Constituents	Description	Examples
Type of subjects	Lexical nouns	Any nouns used as the subject of *be* construction
	Personal pronouns	I, you, he, she, it, we, they
	Demonstrative pronouns	this, that, these, those
	Indefinite pronouns	one, some, all, both, many, no one etc.
	Wh-subjective pronouns	who, which,
Type of subject predicates	Noun phrase	The kernel **is** *the part of the plant of greatest value*.
	Adjective phrase	That **wasn't** *very nice*.
	Prepositional phrase	The houses **are** *in the conservation area.*
	Infinitive *to* clause	The capital **is** *to be provided by the French government*.
	That-clause	But the danger **was** *that the pound would fall further than planned.*
	Wh-clause	That**'s** *why I bought the refill*.
	Indefinite noun phrase	There's *a man* sitting in the corner.
	Definite noun phrase	There **are** *around 6000 accidents* in the kitchen of Northern Ireland home every year.
Post-*BE* verb	Transitive verbs	verbs require some type of objects (take, lick, eat,)
	Unergative verbs	intransitive verbs with volitional acts like (dance, shout, work, cry, laugh)
	Unaccusative verbs	intransitive verbs with non-volitional acts (melt, happen, sink, appear, break, fall)
Auxiliaries	*have*	has, have, had
	Modals	can, may, shall, will, could, might, should, would
	Semi-modal	have to, ought to, be going to, be supposed to, have got to, used to
Intensifiers	Degree adverbs	very, always, so, extremely, highly, greatly
	Negation	not

3.1.4.3 Analytical Parameters for Ungrammatical Use of *BE*

The analytical parameters for the ungrammatical use of *BE* are set based on the types of errors already attested in previous studies. Several major types of errors of both copula and auxiliary *BE* have been documented in previous studies and they include omission, overgeneration, agreement and tense errors. These error categories can be traced back to the surface structure taxonomy proposed by Dulay, Burt and Krashen (1982). Dulay et al. (1982) categorised learner errors to four major categories, namely omission, addition, misformation and misordering. Omission errors reported in this study bears similar characteristics to the omission errors proposed by Dulay et al. (1982), whereby they involve dropping a function word in obligatory context. Overgeneration, which according to past studies involves the insertion of *BE* where it is not required (e.g. *was came*) falls under the addition error category, while both agreement and tense errors belong to misformation category. This study shall categorise the errors according to the classification used in previous studies (omission, overgeneration, agreement, tense)

mainly because the reference to the errors are more specific to errors in the use of *BE* and the terms are more commonly used in more recent studies.

Omission of *BE* has been recorded as one of the major types of errors among Malaysian ESL learners. Maros et al. (2007), Wee (2009) and Wee, Sim & Kamaruzam (2010) found that omissions of copula and auxiliary *BE* to be a very common type of errors found in the written works of these learners. This type of errors was also found in the written data of ESL learners of L1-Chinese (Lee & Huang, 2004), L1-Sinhala (Herat, 2005), L1-Persian (Kafipour & Khojasteh, 2012) and L1-Arabic (Muneera & Wong, 2011; Murad & Khalil, 2015).

ESL learners are also found to produce *BE* overgeneration, where *BE* is inserted in non-obligatory context and used with a lexical verb to produce constructions such as '*The queen is come*' (Lee & Huang, 2004). Overgeneration type errors are also common in the language data of Malaysian ESL learners (Arshad & Hawanum, 2010; Maros, et al., 2007; Wee, 2009; Wee, Sim & Kamaruzam, 2010). The same construction was also found in the language samples of Russian (Ionin & Wexler, 2001; 2002), Chinese (Ju, 2000; Balcom, 1997; Lee & Huang, 2004; Yip, 1994), Spanish (Fleta, 2003; Oshita, 2000), Japanese (Hirakawa, 2006; Oshita, 2000) and Korean (Oshita, 2000; Park & Lakshmanan, 2007) learners. It is important to highlight that in the studies by Oshita (2000), Ju (2000), Balcom (1997), Yip (1994) and Park and Lakshmanan (2007), *BE* insertion in non-obligatory context is termed overpassivisation; a term employed by the researchers to describe passive-like construction involving specifically unaccusative verbs (e.g. *was happened*, *was sink*). In the present study, insertion of *BE* carries similar characteristics of those found in the study by Ionin and Wexler (2001), when *BE* is added as a mechanism to mark tense and/or agreement feature, thus, adopts the term overgeneration of *BE* as it is used by Ionin and Wexler (2001).

Other types of errors found in the written data of Malaysian learners are agreement and tense errors. Maros et al. (2007) and Nor Hashimah, Norsimah and Kesumawati (2008) found that the L1-Malay learners involved in their study had the tendency to produce both agreement and tense errors. Arshad and Hawanum (2010), who analysed the written data of Malaysian learners from three L1 backgrounds; Malay, Chinese and Tamil, found tense errors to be among the major types of errors produced by their subjects. Another pair of Malaysian researchers, Siti Hamin and Mohd Mustafa (2010) in their investigation of Malaysian learners' agreement errors, reported occurrences of agreement errors involving both copula and auxiliary *BE.*

Based on the learner errors already attested in previous studies, the analytical parameters for the ungrammatical use of *BE* for this study are set as summarised in Table 3.13.

Table 3.13: Analytical Parameters for Ungrammatical Use of *BE*

Ungrammatical Use	Description	Examples
Omission	*BE* is omitted in an obligatory context	*One of the boys Ø playing football with his friend.* (Muneera & Wong, 2011)
Agreement	*BE* does not agree with the subject	*Their students **is** in good health.* (Siti Hamin & Mohd Mustafa, 2010)
Tense	wrong tense	*In a kingdom, there **is** [was] a very beautiful princess.*(Maros, et al., 2007)
Overgeneration	*BE* inserted before a lexical verb in non-obligatory context	*The nurse **was bandaged** her leg.* (Wee, 2009)

3.1.4.4 Data Coding Process

This section describes in detail the data coding process involved in the coding of LOCNESS and L1-Malay learner sub-corpus.

3.1.4.4.1 Coding LOCNESS

Coding the forms, functions and pre-*BE* /post-*BE* constituents for LOCNESS involves a rather straight forward process. CLAWS C5 tagsets developed by University Centre for Computer Corpus Research on Language, Lancaster University (Garside & Smith,

1997) was utilised to POS tag LOCNESS. The C5 tagsets have only over 60 tags. It is comparatively simpler and smaller than C6 and C7, which have over 160 tags. Nevertheless, it contains sufficient tagsets to code the *BE* forms in LOCNESS to be used for the purpose of this study. Once the NS learner corpus was tagged, the *BE* forms and the pre-*BE* /post-*BE* constituents can be easily extracted using WordSmith Tools. Table 3.14 displays the tagsets for *BE* assigned by C5 tagsets (the complete tagsets are available in Appendix B).

Table 3.14: C5 Tagsets for *BE*

Tagset	Description
VBB	the "base forms" of the verb "BE" (except the infinitive), i.e. AM, ARE
VBD	past form of the verb "BE", i.e. WAS, WERE
VBG	-*ing* form of the verb "BE", i.e. BEING
VBI	infinitive of the verb "BE"
VBN	past participle of the verb "BE", i.e. BEEN
VBZ	-*s* form of the verb "BE", i.e. IS, 'S

3.1.4.4.2 Coding of L1-Malay Learner Sub-Corpus

This section describes in detail the process involved in coding the forms and functions of *BE* in the L1-Malay learner sub-corpus.

A. Development of tagsets

Step 1: Setting the criteria

The first step in the coding process is to develop a comprehensive tagsets to enable the data to be analysed according to the research objectives of the study. In doing so, the tagsets will have to be developed based on the analyses parameters already specified; forms of *BE* (Table 3.9), functions of finite *BE* (Table 3.10), functions of non-finite *BE* (Table 3.11), pre-*BE* and post-*BE* constituents (Table 3.12) and the ungrammatical use of *BE* (Table 3.13). These parameters can be summarised as the followings:

1. Forms: Finite *BE* forms vs. non-finite *BE* forms.
2. Functions of *BE*: copula, auxiliary, interrogative, negative operator, existential *there*, *it*-cleft.
3. Types of subjects: nouns and pronouns.

4. Types of predicates: nominal, adjectival, prepositional and clausal predicates.
5. Post-*BE* verbs: forms-inflected and uninflected; classes-transitive, unergative and unaccusative.
6. Other pre-*BE* and post-*BE* constituents: auxiliaries and intensifiers.
7. Ungrammatical use: omission, overgeneration, tense and agreement.

Step 2: Selecting the tag syntax

The tags were developed using the SGML (Standard Generalised Mark-up Language) syntax, which involves the use of a pair of balanced angled brackets; <….> was used for a start tag and </…> for the end tag (McEnery & Wilson, 2001). The end tag contains a slash character preceding the annotation strings to mark the end of the annotation. As an example the word "Censorship" in the following sentence is tagged <NP>*Censorship*</NP>, where the tag "NP" is used to denote noun phrase.

A0004

```
<NP>Censorship</NP>  <FAuxPas>is</FAuxPas>  <Vt-
en>defined</Vt-en>  <Cnj>as</Cnj>  <NP>the  act,
process</NP>  <Conj>or</Conj>  <NP>practice  of
censoring</NP> <Conj>or</Conj>  <NP>banning</NP>.
```

The advantages of using SGML syntax are, "it is simple, clear, formally rigorous and already recognised as an international standard" (McEnery & Wilson, 2001, p. 35). Besides that, it can be retrieved easily using the data retrieval software adopted for this study; WordSmith Tools version 5 (Scott, 2008). The syntax also enables the structure of the text to be clearly identified and it is also found to be very useful for web-based pedagogical tools or databases as hypertexts (Izumi, Uchimoto, & Isahara, 2005).

Step 3: Developing the tagsets

The tagsets are developed according to these divisions; (a) tags denoting the finiteness and function of the *BE*, (b) tags for pre-*BE* and post-*BE* constituents and (c) tags for the ungrammatical use of *BE*.

a) Tags for the functions of *BE*

The tagsets designed for this purpose include two main information; the finiteness and the function of *BE*. To indicate the finiteness of the verb the initial *F* (denotes finite) and *NF* (denotes non-finite) are adopted. These initials are attached with the main tagset denoting the function. The tagsets for denoting functions are made up of three initials of the function, with the first initial capitalised for example -*Cop* to denote copula *BE*. *BE* is divided into two main functions, copular and auxiliary, which are tagged <FCop> and <FAux> respectively. Since auxiliary *BE* can be used as either a progressive aspect marker or in the formation of passives, this category is further divided to auxiliary progressive and auxiliary passive which are then respectively assigned <FAuxPro> and <FAuxPas> tags. Another information attached to the *BE* tagset is the verb status as a negative operator, which is assigned the initials *Neg*. As an example, a copula *BE* that also functions as a negative operator is assigned <FCopNeg> tag. As for other syntactic information namely existential *there*, *it*-cleft, and interrogative operator, each is assigned <Ext>, <It> and <Qs> tag respectively. These tags will be placed separately at the beginning of a sentence containing *BE* as shown in the following sample sentences:

(c) **<Qs>**<FCop>Is</FCop> <PPN>it</PPN> <NP>the line of old song</NP> <AP>right</AP>**</Qs>**?

(c) ***<Ext>There*** <FCop>are</FCop> <NP>a lot of reasons</NP> <ThtC>that made me agree with the line of old song</ThtC>**</Ext>**.

Table 3.15 below displays the complete set of tags used to code the functions of *BE*.

Table 3.15: Tagsets Denoting Functions of *BE*

Tagset	Description
<FCop>	finite copula
<FCopNeg>	finite copula *BE* forms as negative operator
<FAux>	finite auxiliary *BE*
<FAuxNeg>	finite auxiliary *BE* forms as negative operator
<FAuxPro>	finite auxiliary *BE* in progressive aspect
<FAuxProNeg>	finite auxiliary *BE* in progressive aspect as negative operator
<FAuxPas>	finite auxiliary *BE* in passive voice
<FAuxPasNeg>	finite auxiliary *BE* in passive voice as negative operator
<NFProB>	non-finite progressive *BE* form: *being*
<NFPasB>	non-finite passive *BE* form: *been*
<NFB>	non-finite infinitive *BE* form: *be*
<Qs>	*BE* as interrogative operator
<It>	expletive '*it'* construction or '*it-cleft'* construction
<Ext>	existential *'there'* construction

b) Tags for pre-*BE* and post-*BE* constituents

The pre-*BE* and post-*BE* constituents include all constituents that occur before and after *BE*. They include the subjects, the post-*BE* verbs (for auxiliary *BE*), the subject predicates (for copula *BE*) and the presence of auxiliaries and intensifiers.

Type of subjects: The subjects are categorised into lexical noun subjects and pronouns. The lexical noun subjects are simply assigned the tag <NP>. They can constitute a single-word noun or a noun phrase. As for pronouns, they are further categorised into several classes, whereby each class is assigned a different tagset as summarized in Table 3.16.

Table 3.16: Tagsets for Types of Subjects

Tagset	Description
<NP>	lexical noun/ noun phrase
<PPN>	personal pronouns (I, you, he, she, it, we, they)
<DPN>	demonstrative pronouns (this, that, these, those)
<IPN>	indefinite pronouns (one, some, all, both, many, no one etc)
<QPN>	subjective wh-pronoun (who)

Form and class of post-*BE* verbs: The verbs following *BE* are tagged according to their morphological, syntactic and semantic features. Different tags are used to represent the various inflectional forms of *BE* (third person singular *–s*, past

tense/past participle *–ed*, progressive *–ing*) and the different types of verb classes (transitives, unergatives and unaccusatives). Table 3.17 summarises the tagsets assigned for coding of post-*BE* verbs.

Table 3.17: Tagsets for Post-*BE* Verbs

	Transitive	**Unaccusative**	**Unergative**
Base forms	<Vt>	<Uac>	<Uer>
3rd person singular forms	<Vt-s>	<Uac-s>	<Uer-s>
Past tense	<Vt-ed>	<Uac-ed>	<Uer-ed>
Past participle forms	<Vt-en>	<Uac-en>	<Uac-en>
Present participle forms	<Vt-ing>	<Uac-ing>	<Uer-ing>

Type of subject predicates: Copula *BE* takes either phrasal or clausal predicates. The phrasal predicates that often occur as the complements of copula *BE* are noun phrase, adjective phrase or prepositional phrase. As for clausal predicates, infinitive *to* clause, *that*-clause and *wh*-clause often take the position as copula *BE* complements. The tagsets for these predicates are displayed in the Table 3.18:

Table 3.18: Tagsets for Subject Predicates

Tagset	**Description**
<NP>	Noun Phrase
<AP>	Adjective Phrase
<PP>	Prepositional Phrase
<InfC>	Infinitive *to* Clause
<ThtC>	*That*-clause
<WhC>	*Wh*-clause

Auxiliaries and intensifiers: Other constituents that are coded and analysed for this study include the auxiliary *have*, modal auxiliaries and intensifiers before and after *BE*. Auxiliary *have* includes *have, has* and *had*, modal auxiliaries include all modals including negative modals like *cannot* and semi-modals like *be going to, have got to*, and finally the intensifiers include negation *not* and degree adverbs such as *very*, *so* and *always*. Negation *not* and degree adverbs

are assigned the same <Adv> tag, following their parts of speech. The full set of the tags used for annotating these constituents are displayed in the Table 3.19.

Table 3.19: Tagsets for Auxiliaries and Intensifiers

Tagset	Description
<Mod>	Modals (can, will, would, should)
<ModNeg>	Modals in negative forms (cannot, won't, can't)
<SMod>	Semi-modals (be going, to, have got to)
<Adv>	Adverb (used for negation *not* and degree adverbs e.g. *always, so, very etc.*)
<VHv>	have
<VHs>	has
<VHd>	had

(c) Tags for ungrammatical use of *BE*

The ungrammatical use of *BE* include errors in tense, agreement, omission and overgeneration. The tags for tense, agreement, and overgeneration are derived from three initials of the misused item with the first initial capitalised, for example <Tns> for tense, <Agr> for agreement and <Ovg> for overgeneration. As for omission, the tags assigned to this category consist not only information of the type of misuse, but also the function of the omitted *BE*. As an example, the omission of copula *BE* is assigned <NCop> in which the letter N stands for *null* and initials Cop for *copula*. Table 3.20 displays the complete tagsets developed for coding the ungrammatical use of *BE*.

Table 3.20: Tagsets for Ungrammatical Use of *BE*

Tagset	Description
<Tns>	Tense
<Agr>	Agreement
<Ovg>	Overgeneration of *BE*
<NCop>	Omission of copula *BE*
<NAux>	Omission of auxiliary *BE*
<NProB>	Omission of progressive *BE* form; *being*
<NPasB>	Omission of passive *BE* form; *been*
<NB>	Omission of infinitive *BE*

B. Manual coding process

This section describes in detail the steps involved in the manual coding process.

Step 1: Software application for manual tagging

Before proceeding in detail with the coding process, it is necessary to explain briefly the software application specifically developed for the manual tagging process. The application, which is named Malaysian Linguistic Tagging Tool (MLTT), was developed using NetBeans, a free and open source integrated development environment (IDE). Among its key features include the ability to manually annotate corpus, customise infinite number of tagsets, add and delete tagsets and calculate the frequency of tagsets already applied. MLTT enables tagstes developed for the coding process to be uploaded onto the application and researcher can manually code the selected linguistic features by simply choosing the tagset and clicking on the item to be coded. Once all the *BE* forms in an essay document have been identified and coded, the document is then saved as a txt.file. Figure 3.1 displays the user interface of MLTT.

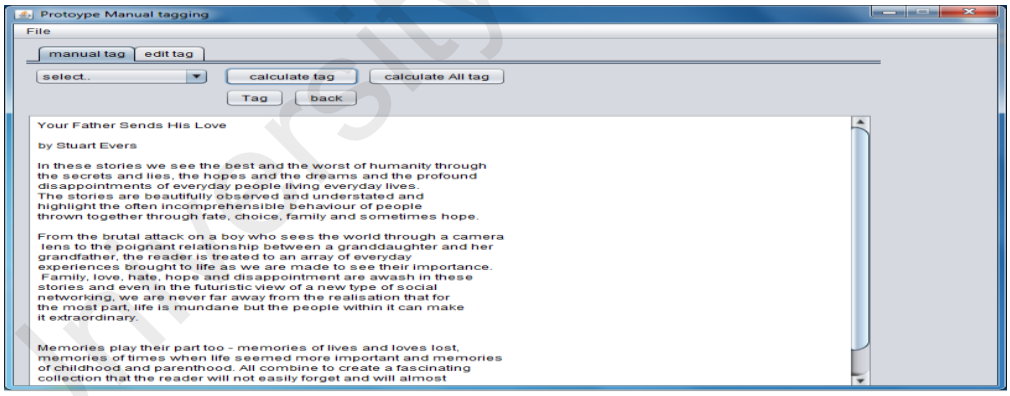


Figure 3.1: The User Interface of MLTT

Step 2: Concordancing all *BE* forms

With the use of WordSmith Tools version 5 (Scott, 2008), all the tokens of *BE* are retrieved from the learner data. They include the various forms and contractions as summarised in Table 3.9. The output for all the *BE* tokens are then stored in separate

files to be used as a checklist that is referred to in ensuring that all *BE* in the L1-Malay learner sub-corpus is coded.

Step 3: Coding *BE* with functions

Tagging the functions entails several other sub-procedures that have to be administered concurrently. They include; (1) classifying the tokens into different functions (e.g. auxiliary, copular) (2) coding for specific syntactic functions, (3) coding pre-*BE* and post-*BE* constituents and (4) coding for ungrammatical use of *BE*.

(1) Classifying *BE* into different functions

In order to decide if a finite *BE* functions as a main verb or an auxiliary is to examine its structure. If it is followed by a main verb in the form of *BE + verb*, it is then tagged as an auxiliary <FAux>. If there is no other main verb present and *BE* itself functions as the main verb, which is used to link the subject to its complement then it is tagged as a main verb or copular <FCop>.

When dealing with learner language data, the coding process is not, however, as straight forward as described above. Complications may arise when handling ungrammatical structure for example overgeneration construction as in *was happened* which is syntactically identical to passive structures (*BE + Ven*) as in (*was called*). In the case when *BE* is found to be inserted/overgenerated in a context where it is non-obligatory, it is then tagged following the procedure described in (4).

Since non-finite *BE* (*be, been, being*) are not considered as main verbs, and generally used in conjunction with auxiliary *have* (have *been*), *modals* (will *be*) and *BE* (is *being*), they are assigned general tagsets; <NFB>, <NFPasB>, <NFPosB> for *be, been* and *being* respectively.

(2) Coding the syntactic functions

Once *BE* is assigned a function for instance as an auxiliary, it has to be concurrently tagged with its syntactic functions. An auxiliary for example can either be used as a marker for progressive aspect or in the construction of passives. If the auxiliary *BE* is followed by V-*ing* and it functions as a marker for progressive aspect, it is then tagged as auxiliary progressive <FAuxPro>, and the auxiliary is tagged <FAuxPas> if it is followed by a past participle and correctly used in the formation of passives.

In addition, *BE* will also be concurrently tagged for other syntactic information namely existential *there*, *it*-cleft, and interrogative operator by using <Ext>, <It> and <Qs> tags respectively. These tags will be placed separately at the beginning of the sentence containing *BE*. In the case when *BE* functions as a negative operator, the information will be attached to the main tag of the verb, for example if *BE* is a negative operator in the progressive aspect, it will be tagged <FAuxProNeg>.

(3) Coding pre-*BE* and post-*BE* constituents

(a) Subjects

The subjects of all *BE* constructions are identified and coded as either lexical nouns or pronouns. Lexical nouns include all proper, common and collective nouns and all the nouns are assigned a common tag <NP>. Pronouns include all personal pronouns, indefinite pronouns, demonstrative pronouns and subjective *wh*-pronouns and each type of pronoun class is assigned a different tag (see Table 3.16). The tags are attached to the nouns or pronouns, for example <PPN>*You*</PPN>, <NP>*Money*</NP>.

(b) Subject Predicates

Copula *BE* can be complemented by either phrasal or clausal subject predicates. Noun phrase, adjective phrase and prepositional phrase are the three

most common types of phrases occurring as the predicates of copula *BE* and each is assigned <NP>, <AP> and <PP> tag respectively. Clausal predicates, namely infinitive *to*-clause, *that*-clause and *wh*-clause, even though are less common, are still used as the subject predicates of copula *BE*. Infinitive *to*-clause is coded with <InfC> tag and *that*-clause with <ThtC> tag and *wh*-clause with <WhC> tag. The subject predicates of copula *BE* are coded following a few steps and these steps are described in detail in Section 3.1.3.4.3.

(c) Post-*BE* Verbs

The verbs preceding *BE* are coded according to their morphological, syntactic and semantic features. Different tags are used to represent the various inflectional forms of *BE* (third person singular *–s*, past tense/past participle *–ed*, progressive *–ing*) and the different types of verb class (transitives, unergatives and unaccusatives). Section 3.1.3.4.4 presents in detail the taxonomy of verbs, the diagnostics used to categorise the verb classes and the process involved in coding the post-*BE* verbs.

(d) Auxiliaries and intensifiers

The study also includes analyses of other pre-*BE* and post-*BE* constituents that might have influenced the grammatical and ungrammatical uses of *BE*. These constituents include auxiliaries and intensifiers that occur before or after *BE*. These constituents are also assigned their corresponding tagsets (refer to Table 3.19).

(4) Coding the ungrammatical use of *BE*

Based on the findings of previous studies the most common types of errors in the use of *BE* are tense, agreement, overgeneration and omission and they are given the tags

<Tns>, <Agr>, <Ovg> and <NCop> respectively. The tags are placed next to the main tag as exemplified in the following sentences

i. In the future, <PPN>it</PPN> ***<Tns><FCop>was</FCop></Tns>*** <AP>dangerous</AP> to our country and community because ....

ii. <PPN>they</PPN> ***<Agr><FAuxPas>is</FAuxPas></Agr>*** <Vt-en>allowed</Vt-en> <InfC>to rejoin society</InfC>.

(a) Tense

Errors in tense (Tns) is identified by first examining the time expressions used, expressions such as *today*, *presently* or *currently* indicate present state, while expressions like *yesterday*, *last year* or *in 1992* signal past state. *BE* must then correspond to the time expressions used. As an example, in the sentence extracted from the learner data *"*This kind of thing* <u>*in the past*</u> *<Tns><FCop>**is**</FCop></Tns> just a dream* <u>*in the past*</u>", the learner used the time expression "<u>*in the past"*</u> twice in the same sentence indicating past reference, which does not correspond with the present copula *is*.

The second method is by looking for any signposts that transmit past/present state such as the use of adverbs like *already* or *just*. If there are no signposts available, the verb has to be examined in its context. The context helps determine if a verb is used in the correct tense.

(b) Agreement

As for agreement errors, the steps taken for coding them is fairly straight forward. First, the verb is examined with its subject. If the subject does not agree with the corresponding *BE* (e.g. a plural form *are* is used with a singular subject as in *money are*) then *BE* is tagged <Agr> to indicate error in agreement.

(c) Overgeneration

Another common error committed by ESL learners is overgeneration of *BE*, borrowing the term used by Ionin and Wexler (2001) for insertion of *BE* before a lexical verb to produce an ungrammatical *BE + V/Ved* construction as in '*is come, was fall* or *was happened'*. Overgenerated *BE* is assigned <Ovg> tag, and the lexical verb is tagged according to its verb class and form. For instance, in an overgeneration consisting of a past tense form of a transitive verb, the verb is then tagged *<Vt-ed>took</Vt-ed>*, where the participle *–ed* that is attached to the main tag denotes past form. The tagsets for the post-*BE* verbs are presented in detail in Table 3.17.

Step 4: Coding *BE* omission

Another vital step in the coding process is to identify and code omission of *BE*. This process is extremely laborious as it requires careful examination of the entire corpus and manually locating and coding the missing *BE*. The following steps explain the process in coding omission of *BE*.

(1) Coding null *BE*

A set of tags are used to code null *BE*, they include <NCop> for copula *BE* omission and <NAuxPro> for auxiliary *BE* omission in progressive aspect and <NAuxPas> for auxiliary *BE* omission in a passive construction. In addition, the tags also include the form of the missing *BE* for example <NCop>*are*</NCop> to indicate the omission of copula *are* or <NAuxPas>*is*<NAuxPas> to annotate missing auxiliary *is* in passive construction. The tags are placed in the position where *BE* should be present as exemplified in the following samples:

i. <PPN>They</PPN> ***<NCop>are</NCop>*** <AP>so excited</AP> <InfC>to gain the money</InfC>

ii. …<PPN>he</PPN> ***<NAuxPro>is</NAuxPro>*** AdvP>always</AdvP> <Vt-ing>wasting</Vt-ing> <NP>his money</NP> <PP>on buying somethings</PP>

(2) Coding pre-null and post-null *BE* constituents

Coding of pre-/post-null *BE* constituents is administered immediately after coding the null *BE*. This involves careful examination of the types of subjects, predicates, form and class of post-null *BE* or any other constituents before and after the null *BE*.

3.1.3.4.3 Coding Subject Predicates

Before describing the process involved in coding the predicates following copula *BE*, there is the need to first discuss the different types of subject predicates. Copula *BE* generally takes two major types of predicates; phrasal predicates and clausal predicates.

A. Type of subject predicates

1. Phrasal Predicates

The phrasal predicates can be in the form of a noun phrase, adjective phrase or prepositional phrase. The followings are the descriptions of these subject predicates:

a. Noun Phrase

According to Biber, et al. (1999) copula *BE* is most frequently complemented by a noun phrase. The noun phrase has two common functions, to characterise and identify the subject noun phrase. The functions of a noun phrase complement are exemplified below:

i. Characterising
*Oh, my dad **was** a great guy, too.*

ii. Identifying
*The kernel **is** the part of the plant of greatest value.*

Biber et al. (2002, p. 142)

b. Adjective Phrase

Adjectivals also occur frequently with copula *BE*. In general they are used to characterise the subject of a sentence. They can be preceded without complements as in (iii) and with either phrasal or clausal complement as in (iv) and (v). Prepositional phrase, *to*-infinitive clause and *that*-clause are the common complements for adjective predicates. The followings are samples of adjective predicates that are used as complements to copula *BE*:

iii. Without complements
*That **wasn't** very nice.*
*It **was** funny though.*

iv. With phrasal complements
*Well you'**re** good [at remembering number].*
*That**'s** nice [of you].*

v. With clausal complements
*I **am** sure [the warm affinities between Scots and Jews arise out of appreciation of herrings.]*

Biber et al. (2002, p. 200)

c. Prepositional Phrase

Prepositional predicates are less common as complements of copula *BE* (Biber et al., 1999). They are generally used to describe a characteristic of the subject and used as adverbial to express position and direction as exemplified below:

vi. Describing Characteristic of Subject
*Umuofia **was** in a festival mood.*
*The resistive voltage drop **is** in phase with the current.*

vii. Adverbial
*I wish you **were** at the shack with me last night*
*The houses **are** in the conservation area.*

Biber et al. (2002, p. 142)

2. Clausal Predicates

Clausal predicates, namely *to*-infinitive clause, *that*-clause and *wh*-clause are sometimes used as complements to copula *BE*. The followings are examples of the use of these predicates as copula *BE* complements:

viii. *to*-infinitive clause
*The capital **is** to be provided by the French government.*

ix. *that*-clause
*But the danger **was** that the pound would fall further than planned.*

x. *wh*-clause
That**'s** *why I bought the refill*

Biber et al. (2002, p. 142)

B. Coding of subject predicates for this study

This section presents the steps taken to code the subject predicates for this study.

Step 1: Identifying the type of predicates

This process is administered concurrently with the coding of copula *BE*. Once *BE* is identified and coded as a copular <FCop>, the next step is to examine the subject predicate and determine if it is a phrase or a clause. This is done by identifying the presence of verbs in the predicate; phrases do not contain verbs, which is the opposite of clauses.

Step 2: Categorising and coding the predicates

The next step is to categorise the phrases or clauses into their respective categories. For a phrase, the key indicator is the head of the phrase. If a phrase is headed by a noun (*He is [my brother]*) it is then categorised as a noun phrase. Similarly an adjective phrase is headed by an adjective (*The table is [too small]*) and a prepositional phrase by a preposition (*The book is [on the table]*). These phrases are assigned <NP>, <AP> and <PP> tagsets for noun phrase, adjective phrase and prepositional phrase respectively. As for clauses, the head of the clause is also used as the indicator. A *to*-infinitive clause

is headed by infinitive *to* (*The file is [to be kept in the second left drawer]*) and *that*-clause by the pronoun *that* (*The reason for his bad behaviour is [that he is not happy with the school condition]*) and *wh*-clause by a *wh*-word (*That's [why he didn't submit his assignment]*). The infinitive-*to* clause is assigned <InfC> tag, *that*-clause is tagged <ThtC> and wh-clause is tagged <WhC>.

3.1.3.4.4 Coding Class of Post-BE Verbs

Previous studies (Ju, 2000; Oshita, 2000; Yip, 1994) have highlighted the influence of post-*BE* verbs on the constructions of *BE* by L2 learners. They reported a consistent pattern of *BE* inserted before unaccusative verbs (*BE* + *unaccusative*) to produce ill-formed constructions such as *was happened*. The findings from these studies suggest that *BE* constructions in the L2 learner data are sensitive to the class of post-*BE* verbs. In determining the extent the class of post-*BE* verbs influences the use of *BE* in the learner corpus of this study, a detailed analysis of the class of the post-*BE* verbs in the L1-Malay learner sub-corpus has to be administered. The following section presents how the different verbs are classified and coded. Prior to that, it is necessary to first discuss the taxonomy of verbs and the diagnostics for distinguishing the different verb classes.

A. The taxonomy of verbs

Grammarians classify verbs into two main classes; transitive and intransitive. Transitive verbs require some types of objects, either one direct object or two and more object phrases. Intransitive verbs on the contrary, do not require any object. Perlmutter (1978) formulated the Unaccusative Hypothesis (UH) that further distinguishes intransitive verbs into two finer classes; unaccusatives and unergatives. Semantically, unaccusatives are verbs with non-volitional acts like *burn, melt, fall, happen*, while unergative verbs are those entailing volitional acts like *dance, walk, work* etc. Unaccusative verbs can be further sub-divided into alternating unaccusative verbs and non-alternating unaccusative

verbs. Alternating unaccusative verbs such as *break, freeze* and *melt*, have transitive counterparts, whereas non-alternating unaccusative verbs do not. Both the unaccusative and unergative verbs only take one argument.

Working within the Government-Binding approach Burzio (1986 cited in Levin & Rappaport Hovav, 1992) proposed that unaccusative and unergative verbs have syntactic d-structure schematised as the followings:

1. a. Unaccusative Verb :____ [vp V NP]

   b. Unergative Verb : NP [vp V]

As shown in (1) unaccusative verb takes an internal argument, while unergative verb takes an external argument. Unaccusative and unergative verbs also differ in the syntactic mapping of their theta roles. Typically, in a construction involving a transitive verb as in (2a), the agent maps onto the subject position, while the theme maps onto the object position. Unergative verbs (2b) function within the typical syntactic mapping, where the sole argument (agentive) maps onto the subject position. It is different, however, for the unaccusative verbs, as seen in (2c) the sole argument (theme) of an unaccusative verb maps onto the subject position. The examples in (2) illustrate more clearly the syntactic mapping of a transitive, unaccusative and unergative verb:

2. a. Transitive: Mary melted the snow
   Subject Direct Object (Grammatical roles)
   <*Agent*> <*Theme*> (Thematic roles)

   b. Unergative: Mary danced.
   Subject (Grammatical roles)
   <Agent> (Thematic roles)

   c. Unaccusative: The snow melted.
   Subject (Grammatical roles)
   <*Theme*> (Thematic roles)

Park and Lakshmanan (2007, p. 329)

B. Diagnosing unaccusativity and unergitivity in English

There are a number of diagnostic tests that can be used in determining the unaccusative-unergative distinction. They include possibility to appear in resultative constructions, be modified by prenominal perfect/passive participles, occur in middle formation, allow causative and unaccusative alternation and take cognate object (Burzio, 1986; Levin & Rappaport Hovav, 1995).

(a) Resultative construction

It is generalised that resultative phrase may only be predicated by the object of a transitive verb, but never the subject. Levin and Rappaport Hovac (1995) term this condition Direct Object Restriction (DOR). DOR predicts if a verb has no object, it then cannot appear in resultative phrase. As shown in example (3a) and (3b), transitive and unaccusative verbs can appear in resultative phrase, whereas unergative verb (3c) cannot.

3. a. She licked the peanut butter clean. (transitive)
   b. The bottle broke open. (unaccusative)
   c. *Dora shouted hoarse. (unergative)

(b) Prenominal perfect/passive participle modification

The participle of passives and unaccusatives can be converted to adjectival form as can be seen in (4a) and (4b). However, the same conversion is not possible with transitive and unergative verbs, as shown in (4c) and (4d).

4. a. The letter was badly written. (passive)
   a'. The badly ***written*** letter.
   b. The book appeared recently. (unaccusative)
   b'. The recently ***appeared*** book.
   c. The artist painted very much. (transitive)
   c'.*The much-***painted*** artist
   d. The lawyer worked hard. (unergative)
   d'. *The hard-***worked** lawyer.

(c) Middle formation

Middle formation involves a construction in which the logical object occurs in the subject position. Thus, the subjects of middle constructions can only be their internal argument. As such only transitive and unaccusative verbs, which have an internal argument, can occur in middle formation, but not unergative verbs as shown in (5c).

5.	a. Bureaucrats ***bribe*** easily.	(transitive)
	b. The door ***opens*** easily.	(unaccusative)
	c. *He laughs easily	(unergative)

(d) Causative alternation

The d-structure of an unaccusative verb entails that the sole argument as the underlying object. This condition allows for the argument to appear in the object position of a transitive verb or a causative structure as in (6a'). The same alternation, however, is not permitted for unergative verbs as the argument of an unergative verb is the underlying subject, thus, not allowing it to appear in object position of a causative structure. This condition is illustrated in (6b').

6.	a. The window ***broke***.	(unaccusative)
	a'. Pam ***broke*** the window.	(causative)
	b. The baby ***cried***.	(unergative)
	b'. *She ***cried*** the baby.	(causative)

(e) Cognate objects

Burzio (1986) posited that by marking an external argument, unergative verbs have the ability to assign accusative case. This is realised in their ability to take cognate objects as in (7a). Unaccusative verbs on the other hand, do not allow for any cognate object as shown in (7b).

7.	a. Tina ***laughed*** <u>a hearty laugh</u>.	(unergative)
	b. *The ship ***sank*** <u>a slow sinking</u>.	(unaccusative)

The diagnostics to classify unaccusative and unergative verbs identified thus far, are not without flaws. According to Levin and Rappaport Hovav (1995), some diagnostics in particular resultative constructions and causative alternation are more stable than others. For the purpose of this research, all the five diagnostics are adopted, however, heavier reliance is given on the results of the more stable diagnostics (i.e. resultative constructions and causative alternation) to ensure that accurate coding of both unaccusative and unergative verbs can be achieved.

C. Coding of post-*BE* verb class in this study

For the purpose of this study, the post-*BE* verbs analysed include transitives, unaccusatives and unergatives. This section discusses the steps taken to code the three classes of verbs.

Step 1: Transitive versus intransitive verbs

The process of classifying transitive and intransitive verbs is fairly straight forward. The information on the transitivity of verbs is already available in dictionaries and other grammar references. For the purpose of this research, commercially published dictionaries namely *Longman Dictionary of Contemporary English* (2015), Longman English Dictionary Online at http://www.ldoceonline.com/ and grammar references such as *Longman Grammar of Spoken and Written English* (Biber et al., 1999) are used to determine the transitivity of the post-*BE* verbs in the L1-Malay learner sub-corpus.

In cases when the transitivity of a verb is ambiguous, diagnostic tests are applied. The first test is to passivise the verb as only transitive verbs can be passivised. The second test is to insert an adverb between the verb and its complement. If they come from the same node, the adverb insertion would cause the sentence to be ungrammatical. In this case only a transitive verb allows for adverb insertion between the verb and its complement as in *We <u>argued</u> **violently** about the plan* (Zobl, 1989, p. 212).

Step 2: Unaccusative and unergative verbs

The next step is to code the intransitive verbs into unaccusatives and unergatives. In doing so, all the intransitive verbs found were put to the diagnostic tests discussed in the previous section. Refer to Table 3.17 for the tagsets used for coding unaccusative and unergative verbs.

3.1.4.5 Accuracy Of Manual Coding Process

Since the coding of the entire *BE* forms, functions, pre-*BE* and post-*BE* constituents in the L1-Malay learner sub-corpus was done manually, it is therefore necessary to evaluate the accuracy of the coding performance. The performance accuracy is defined by:

$$\frac{\text{Number of tokens correctly coded}}{\text{Number of tokens}}$$

Nagata, Whittaker, Sheinman (2011)

Before presenting the results of the accuracy evaluation, it is necessary to first explain the steps involved in the process. The following steps explain in detail how the evaluation was conducted:

Step 1: Selection of a second coder

The first step was to assign another English language expert to tag a number of essays from the L1-Malay learner sub-corpus. The expert, who was selected to perform this task, has over 20 years of experience teaching English proficiency and ESP courses in a local university. She holds a first degree in Education (TESL Hons.) and a master's degree in Language Studies from Universiti Kebangsaan Malaysia and at the time the study was conducted she holds a position as a senior lecturer at the university she is attached to.

Step 2: Selection of essays to be tagged

It is decided that a total of 36 carefully selected scripts or 10% of the total 366 scripts would be tagged by the second coder. The number of scripts chosen for this purpose is deemed sufficient since manual tagging is a tedious and time consuming process. Considering that the second coder has a full time work commitment, therefore, has very limited time to spend on the coding process, 36 scripts is considered a sufficient number. In addition, the number of scripts chosen is considered quite large since in the study by van Rooy and Schäfer (2002) only 5 scripts were used to evaluate the accuracy of three automatic POS taggers; TOSCA-ICLE, Brill tagger, and CLAWS. The scripts were carefully selected to ensure that all the samples consist most of the items under evaluation. Nevertheless, it is impossible to get an even number of all the possible items to be tagged, but the selection process has enable all the items to be present in the samples.

Step 3: The training

Before the actual tagging process can begin, the coder had to undergo a short training session when she was briefed on (i) the items to be tagged, (ii) the tagsets for the items to be tagged and (iii) the use of MLTT in the tagging process. To ease the process of tagging, the coder is equipped with a guide sheet that consists of all the items to be tagged, an example of the item in context and the tagset for each item (refer to Appendix A). The guide sheet summarises all the analytical parameters and the tagsets to be used. The coder was then briefed on the functions of MLTT and how it is used in tagging the essays. Next, the coder was asked to practise tagging two essays, which was then verified by the researcher. The actual tagging process proceeded once the coder has been adequately informed of and trained for the tagging process.

Step 4: Accuracy evaluation

When all the 36 essays have been tagged, they were then run through WordSmith Tools Version 5 to check for accuracy. This task was undertaken by the researcher. This process involved careful examination and calculation of the coded data. This was done by first running each tagset through the concordancer and the second step was to examine each concordance lines for any inaccuracy. The third step was to calculate the number of inaccurately coded items. The sum of inaccurately coded items was then deducted from the total number of tokens, which then resulted in the number of correctly coded items. The fourth and final step was to calculate the division of the number of correctly coded items to the number of tokens, which then produced the results of the coding accuracy. Table 3.21 summarises the results of the coding accuracy for each of the major tagsets. As shown in Table 3.21, the accuracy reading of the individual item is found to range from 0.875 to 1.000. The overall accuracy is recorded at 0.985 point. In general, the results indicate that the coded data are very stable and reliable.

In terms of the source of inaccuracy, it was found that nearly all of the inaccuracies identified in the data are the results of human error. In most of the cases the items were coded with the wrong tagsets. Nevertheless, as can be seen from Table 3.21 the number of wrongly coded items are very low, showing that the coder are able to code the items consistently using the guide given.

Table 3.21: Accuracy of Manual Coding of L1-Malay Learner Sub-Corpus

		Num of Tokens	Num of tokens correctly coded	Accuracy
Functions of *BE*	<FCop>	421	415	0.985
	<FCopNeg>	35	31	0.886
	<FAuxPro>	455	453	0.995
	<FAuxProNeg>	2	2	1.000
	<FAuxPas>	92	90	0.978
	<FAuxPasNeg>	0	0	0.000
	<NFProB>	0	0	0.000
	<NFPasB>	21	21	1.000
	<NFB>	101	101	1.000
	<Qs>	4	4	1.000
	<It>	21	18	0.857
	<Ext>	56	56	1.000
Subjects	<NP>	414	414	1.000
	<PPN>	165	164	0.993
	<DPN>	50	50	1.000
	<IPN>	7	7	1.000
	<QPN>	23	23	1.000
Post-*BE* Verb Class	<Vt>	24	24	1.000
	<Vt-s>	0	0	1.000
	<Vt-ed>	102	98	0.960
	<Vt-ing>	27	27	1.000
	<Uac>	3	3	1.000
	<Uac-s>	0	0	0.000
	<Uac-ed>	0	0	0.000
	<Uac-ing>	0	0	1.000
	<Uer>	8	8	1.000
	<Uer-s>	0	0	0.929
	<Uer-ed>	4	3	0.750
	<Uer-ing>	15	14	0.933
Subject	<NP>	229	227	0.991
Predicates	<AP>	158	152	0.962
	<PP>	42	42	1.000
	<InfC>	15	13	0.866
	<ThtC>.	8	8	1.000
	<Wh-clause>	10	10	1.000
Ungrammatical Use	<Tns>	8	7	0.875
	<Agr>	0	0	0.000
	<Ovg>	39	36	0.923
	<NCop>	8	8	1.000
	<NAux>	10	9	0.900
Total		2577	2538	0.985

3.1.5 Data Analysis Procedure

This section presents in detail the procedure involves in analysing L1-Malay learner sub-corpus and LOCNESS.

3.1.5.1 Analysis Procedure for L1-Malay Learner Sub-Corpus

Data analysis commenced once the entire learner corpus has been coded. The analysis focuses on three major aspects of the research; (1) the distribution of all forms and functions of *BE* (2) the patterns of grammatical and ungrammatical uses of *BE* and (3) the possible influence of pre-*BE* and post-*BE* constituents on the patterns of the grammatical/ungrammatical use of *BE*.

Step 1: Calculating the frequency of all forms and functions of finite *BE*

In order to obtain the frequency of all the forms and functions *BE* in the learner corpora, the first step is to run the coded data using WordSmith Tools 5. This generates the frequencies of all the grammatical and the ungrammatical forms and functions. When dealing with a large amount of data, it is necessary to have a clear guideline to keep the analysis on the track. Hence, the following list acts as a guideline to help ensure the analyses do not diverge from the objectives and most importantly it also helps to ensure thorough and comprehensive analysis.

1. The tokens of finite and non-finite *BE* forms (e.g. *is, are, was, were, be, been, being*) in the entire corpus;
2. The frequency of *BE* according to functions (e.g. main verb, auxiliary, negative operator, interrogative operator, existential *there* & *it*-cleft);
3. The frequency of *BE* errors; tense, agreement, overgeneration and omission.

The results obtained from these analyses, provide the overall distributional patterns of the various forms and functions of *BE* in the L1-Malay learner sub-corpus. The figures are then compared to the native learner corpus. This comparison enables for analyse the differences and similarities of the use of *BE* between the NS speaker leaners and L1-Malay learners to be conducted. In addition, it also reveals information on which forms

and functions that are more likely to be grammatically and ungrammatically used by the L2 learners.

Step 2: Analysing the linguistic patterns of *BE* and mapping forms to functions;

The next step is to analyse each instance of *BE* with regard to (1) position of *BE* in the sentence, and (2) the syntactic properties of the pre-*BE* and post-*BE* constituents. The linguistic patterns are extracted by analysing the concordance outputs of relevant tags. The following list is used as a guideline for the analysis:

A. When *BE* functions as a copular:
    1. How frequently does copula *BE* occur in the following patterns?
        a. NP/PN + *BE* + NP/AP/PP or
        b. NP/PN + *BE* + InfC/ThtC/WhC.
    2. How frequently is copula *BE* ungrammatically used in the following patterns?
        a. NP/PN + *BE* + NP/AP/PP or
        b. NP/PN + *BE* + InfC/ThtC/WhC.
    3. How frequently is copula *BE* ungrammatically used in the following patterns?
        a. NP/PN + *BE* + Adv/Not/Mod + NP/AP/PP or
        b. NP/PN + *BE* + Adv/Not/Mod + InfC/ThtC/WhC.
    4. How frequently is copula *BE* omitted resulting in the following patterns?
        a. NP/PN + NP/AP/PP or
        b. NP/PN + InfC/ThtC/WhC .
    5. How frequently is copula *BE* omitted resulting in the following patterns?
        a. NP/PN + Adv/Not + NP/AP/PP or
        b. NP/PN + Adv/Not + InfC/ThtC/WhC .

B. When *BE* functions as an auxiliary:
    1. How frequently does auxiliary *BE* occur in the following patterns?
        a. NP/PN + *BE* + Vt.

b. NP/PN + *BE* + Uer.

c. NP/PN + *BE* + Uac.

2. How frequently is auxiliary *BE* ungrammatically used in the following patterns?

a. NP/PN + *BE* + Vt.

b. NP/PN + *BE* + Uer.

c. NP/PN + *BE* + Uac

3. How frequently is auxiliary *BE* ungrammatically used in the following patterns?

a. NP/PN + *BE* + Adv/Not + Vt.

b. NP/PN + *BE* + Adv/Not + Uer.

c. NP/PN + *BE* + Adv/Not + Uac.

4. How frequently is auxiliary *BE* omitted resulting in the following patterns?

a. NP/PN + Vt/Uer/Uac?

6. How frequently is auxiliary *BE* omitted resulting in the following patterns?

a. NP/PN + Adv/Not + Vt/Uer/Uac?

C. When non-finite *BE* forms are used:

1. How frequently does infinitive *be* occur in the following patterns?

a. NP/PN + Mod + *be* + AP/PP,

b. NP/PN + Mod + *be* + Ven + PP,

c. NP/PN + Mod + *be* + Ving + NP/AP/PP?

2. How frequently is infinitive *be* ungrammatically used in the following patterns?

a. NP/PN + Mod + *be* + AP/PP,

b. NP/PN + Mod + *be* + Ven + PP,

c. NP/PN + Mod + *be* + Ving + NP/AP/PP.

3. How frequently is infinitive *be* omitted resulting in the following patterns?

a. NP/PN + Mod + AP/PP,

b. NP/PN + Mod + Ven +PP,

c. NP/PN + Mod + Ving + NP/AP/PP?

4. How frequently does *been* occur in the following patterns?
    a. NP/PN + have + *been* + AP/PP,
    b. NP/PN + have + *been* + Ven + PP,
    c. NP/PN + have + *been* + Ving + NP/AP/PP.
5. How frequently is *been* ungrammatically used in the following patterns?
    a. NP/PN + have + *been* + AP/PP,
    b. NP/PN + have + *been* + Ven + NP/AP/PP,
    c. NP/PN + have + *been* + Ving + NP/AP/PP.
6. How frequently is *been* omitted resulting in the following patterns?
    a. NP/PN + have + NP/AP/PP,
    b. NP/PN + have + Ven + NP/AP/PP,
    c. NP/PN + have + Ving + NP/AP/PP.
7. How frequently does *being* occur in the following patterns?
    a. NP/PN + be + *being* + Ven + AP/PP?
8. How frequently is *being* ungrammatically used in the following patterns?
    a. NP/PN + be + *being* + Ven + AP/PP?
9. How frequently is *being* omitted resulting in the following patterns?
    a. NP/PN + Ven + AP/PP?

Step 3: Focused analysis on distinct ungrammatical use of *BE*; overgeneration and omission of *BE*.

The next step in the analysis procedure is to conduct a detailed examination on the distinct ungrammatical constructions found, namely overgeneration and omission. The investigation takes the following directions:

A. Overgeneration of *BE* before lexical verb (*BE + V*) is analysed according to:
    1. The form of the lexical verbs: *V, V-s, V-ed, V-ing*.
    2. The class of the lexical verbs: transitives, unergatives, unaccusatives.

3. The subjects: noun, pronouns.
4. The subject predicates: noun phrase, adjective phrase, prepositional phrase or complement clauses.
5. The presence of intensifiers and auxiliaries.

B. Omission of *BE* is analysed according to:

1. Functions: copular and auxiliary.
2. Finiteness: finite (e.g. is, *are, was, were*) and non-finite (e.g. *be, been, being*).
3. The subjects: noun, pronouns.
6. The subject predicates: noun phrase, adjective phrase, prepositional phrase or complement clauses.
7. The class of the lexical verbs: transitives, unergatives, unaccusatives.
4. The presence of intensifiers and auxiliaries.

3.1.5.2 Analysis Procedure for LOCNESS

LOCNESS is to be analysed to obtain (i) the frequency of all of the finite and non-finite *BE* forms and (ii) the frequency the two main functions of *BE*: copular and auxiliary. In general the analysis procedure for LOCNESS follows these steps:

Step 1: Dividing LOCNESS into two sub-corpora.

Before LOCNESS can be analysed for the overall patterns of the use of *BE*, it will be first divided into two sub-corpora; American (Ame.) learner sub-corpus and British (Bri.) learner sub-corpus, representing the American variety and British variety respectively. The two sub-corpora are analysed separately in order to determine the similarities and differences in the distribution of *BE* in each variety.

Step 2: Calculating the frequency of all forms of finite and non-finite *BE.*

This analysis is administered to obtain the overall distribution of *BE* across both sub-corpora. The same analytical parameters for all finite and non-finite *BE* forms for L1-Malay learner sub-corpus (refer to Table 3.14) will be utilised for the analysis. The results for this analysis reveals firstly the similarities and differences in the distribution of *BE* across the two NS learner sub-corpora and secondly the similarities and differences in the distributional patterns between the NS learner sub-corpora and NNS learner corpus.

Step 3: Calculating the frequency of *BE* according to two main functions (copular and auxiliary).

The second analysis is to obtain the patterns of use of *BE* in terms of its major functions and for this purpose the analytical parameters for the functions of *BE* in Table 3.15 will be referred to. The analysis of the functions of *BE* in LOCNESS is restricted to only two major functions; copular and auxiliary, since the purpose of this analysis is to obtain the general patterns of *BE* use by the NS learners that can be compared to that of the NNS learners. The comparison within the NS sub-corpora that is between the American and British sub-corpora is also administered to see if there are differences in the overall patterns of *BE* in the sub-corpora.

3.1.5.3 Comparative Analysis Procedure for L1-Malay Learner Sub-Corpus and LOCNESS

As shown in Figure 3.2 the comparative analysis in this study involves the comparison between the use of *BE* within LOCNESS corpus (intra-corpus analysis) and between LOCNESS and the L1-Malay learner sub-corpus (inter-corpora analysis). This section describes the procedure involved in these comparisons.

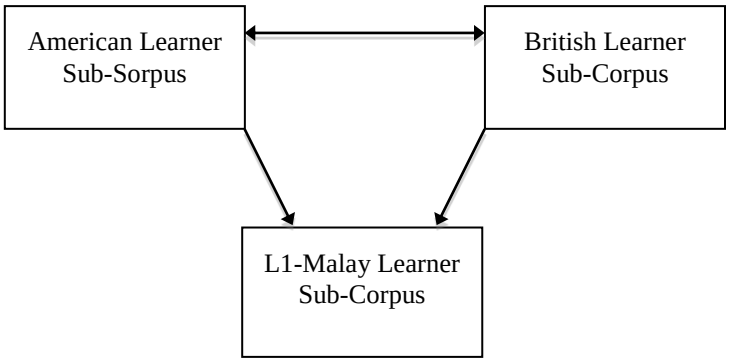


Figure 3.2: Inter- and Intra-Corpora Comparative Analysis

Step 1: Intra-Corpus Analysis (American and British Learner Sub-Corpora)

Before the comparison between the control corpus (LOCNESS) and the NNS (L1-Malay learner sub-corpus) can be administered, it is important to first compare the patterns in the use of *BE* between the American and British learners. The analysis is administered to reveal differences in the patterns of use. If the analysis revealed significant differences between the two sub-corpora, then each sub-corpus will be compared to the L1-Malay learner sub-corpus separately as shown Figure 3.12. In other words the sub-corpora are then treated as two different varieties of the NS English, therefore, the results of the patterns of use in the L1-Malay learner sub-corpus have to be compared to the results obtained from each of the LOCNESS sub-corpora. The comparison will be conducted on the patterns of all the forms of finite and non-finite *BE*, as well the patterns of the use of *BE* according to its major functions (copular and auxiliary).

This study assumes that both the American and British learners would employ considerable frequency of *BE* in their writings and there would be not much difference in the patterns of use of *BE* between American and British learners. This assumption is made based on findings of previous studies in particular by Staples and Reppen (2016)

in their analysis of the choice of verbs used by learner writers in expressing stance in long arguments. They found that L1 English learners notably used more *BE* controlling complement clauses, which has allowed the L1 writers to express stance less overtly and this method is considered more effective than using conversational verbs such as *think.* Jarvis, Grant, Bikowski and Ferris (2003) also noted high positive score of stative *BE* in two of the six clusters used to categorize highly rated learner essays. The clusters were also characterised by having high positive mean scores for text length, demonstratives, nominalizations and adverbial subordination and high positive mean scores for text length, diversity of vocabulary, and passives, all of which are essential features in academic prose. These studies highlight a very important function of *BE* in academic writing, that clauses with *BE* as the main verb despite their simplicity perform important functions in academic prose. Biber et al. (1999) also highlighted that *BE* is a very common verb in academic prose, where it is used for characterising and identifying the subject NP. Based on these findings it can be assumed that *BE* would be used frequently in both NS learner corpora, as both contain mainly argumentative essays that would require the use of *BE* especially in stance making and characterising and identifying the subject NP.

Step 2: Inter-Corpora Analysis (LOCNESS and L1-Malay learner sub-corpus)

This step involves the comparison between the results of the patterns of use of *BE* according to forms and functions obtained in the American learner sub-corpus with the results obtained from L1-Malay learner sub-corpus and between the results from the British learner sub-corpus to the results from L1-Malay learner sub-corpus.

The aim of these comparative analyses is to compare the overall patterns of the use of *BE* by the L1-Malay learners against the patterns of use by the American and British learners respectively. The main focus of the analyses is on the differences and similarities in the use of *BE* by these learners. The analyses, therefore, would only

comprise of analyses of the distributional patterns of the all the finite and non-finite forms of *BE* and the patterns of use of copula and auxiliary *BE*. The results from these analyses would then reveal the similarities and differences in the use of *BE* by L1-Malay learners with that of the American and British learners.

The study assumes that there would be very significant differences in the patterns of use of *BE* between the L1-Malay ESL learners and both the NS learners (American and British). Clauses with *BE* as the main verb are regarded as syntactically simple (Hinkel, 2002, 2003) and often associated with structures used in conversation. Hinkel (2002, 2003) reported that ESL learners tend to overuse *BE*-copula clauses in their writings and that these clauses have characteristics of spoken rather than written discourse. Most often the constructions involve the use of copula *BE* in simple propositions made up mostly of a subject and an adjective predicate. Clauses with *BE* as the main verb are relatively simpler than those with verbs that have higher semantic and lexical content (Biber et al., 1999) and for this reason it is anticipated that L1-Malay ESL learners especially those with lower proficiency would rely more heavily on simple *BE* constructions. Therefore, it is anticipated that there would be higher frequency in the use of *BE* by the L1-Malay ESL learners than the NS learners.

3.2 Textual Analysis

In order to supplement the findings of the quantitative analyses and to discover how the learners use *BE*, the data have to undergo a set of qualitative analysis, which involves textual analysis of the all the grammatical and ungrammatical uses of *BE*. The qualitative analysis basically uses the same parameters set for the quantitative analysis with more focus given on how *BE* is used by the learners. It shall provide evidence of how *BE* is actually realised for example in *NP+BE+AP* or *NP+BE+NP* structure, so that evaluation on the influence of syntactic environments can be determined more

accurately. According to Sinclair (1991) introspection is better used "in evaluating evidence rather than creating it" (p. 39), emphasising that evidence of use in context is more valuable than samples derive from researcher's intuition since they are more objective with no prejudices and preferences from the researcher.

Many corpus-based studies have adopted this approach to analysing corpus data and some that have been reviewed in this study include Abdullah and Noor (2013) on verb-collocations, Aziz, Jin and Nordin (2016) on metadiscourse and Mohd Don and Sriniwass (2017) on conjunctive adjuncts. In all these studies the qualitative data are used in providing exemplification of how grammatical aspects are used by the learners and are also used in explaining the aggregate data. The present study emulates this approach to integrating qualitative and quantitative analyses, whereby the qualitative data shall supply the exemplication of how *BE* is used in the context for each of the patterns of use unraveled by the quantitative analysis. The use of *BE* in context will also enable the researcher to determine more precisely the extent of the influence of the syntactic environments on the use of *BE*. The analysis shall go a step further by examining also the types of clauses in which *BE* is more likely to occur in order to provide more insights into the actual realisation of *BE* in the L1-Malay ESL learner data.

The qualitative analysis in the present study is divided into three major activities:

1. Extracting the grammatical and ungrammatical uses of *BE* in order to examine the patterns of both the grammatical and ungrammatical uses.
2. Investigating the use of *BE* in relation to the syntactic environments; examining the possible influence of the constituents before and after *BE*.

3. Examining the use of *BE* in relation to syntactic complexity; discovering the types of sentences (simple, compound or complex) and clauses in which *BE* occurs.

3.3 Analysis Framework of the Study

This section displays the diagram that sums up the analysis framework employed in this study. The diagram lays out the entire process involved in the analysis procedure. The process begins with the selection of linguistic features to be investigated i.e. *BE*. The second step is the selection of corpus data, MACLE (Knowles et al., 2006) is chosen to represent the ESL learners' language sample, while LOCNESS (Granger, 1993) is selected as the reference corpus, representing the native learner language sample. The third step is setting the analytical parameters of *BE*, these parameters identify the forms, function, pre-*BE* and post-*BE* constituents and the possible errors involving *BE*. The analytical parameters are set based on two main sources, namely the *Longman Grammar of Spoken and Written English* (Biber et al., 1999) and findings from previous studies. The fourth step involves selection of the data coding tool. CLAWS4 POS tagger is chosen to code LOCNESS, while MACLE is manually coded. The manual coding process entails three major steps; the development of the manual tagsets, manual coding process and administration of accuracy test for the already coded texts. The fifth step is the selection of computerised concordance for the analysis purposes. WordSmith Tools Version 5 is selected to perform the analyses for this study. Both MACLE and LOCNESS are analysed using WordSmith Tools Version 5. The final step for the corpus-based method involves the analysis of *BE* as specified in the analytical parameters.

As for the textual analysis, it requires for the grammatical and ungrammatical uses of *BE* to be extracted and analysed to uncover how *BE* is used by the learners. The instances of *BE* are then analysed in relation to the syntactic environments in order to

examine the possible influence of the surrounding constituents on the use of *BE*. The last step in the textual analysis is to examine the syntactic complexity in which *BE* occurs. Figure 3.3 is the diagram summarising the corpus-based analysis framework adopted for this study:

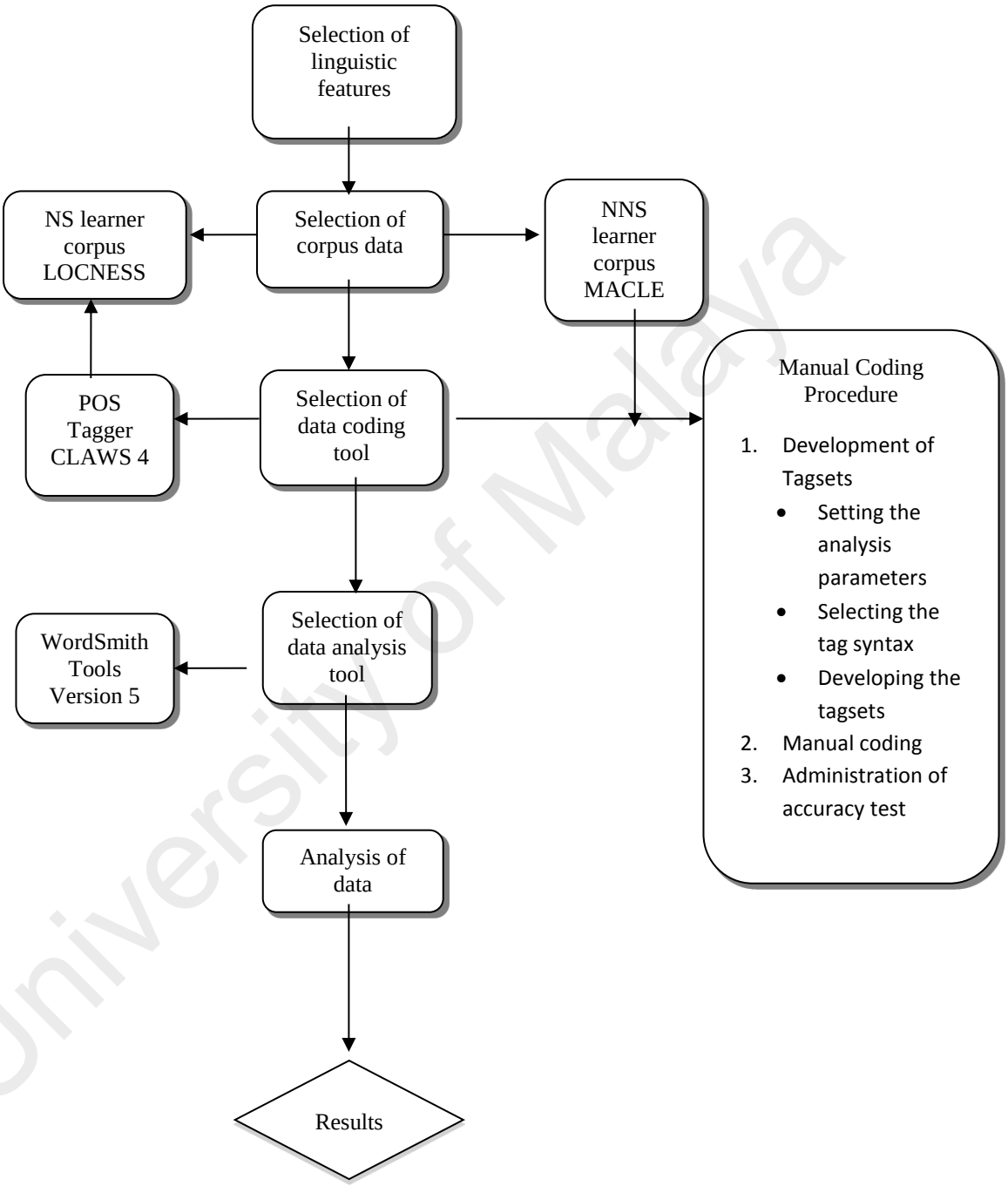


Figure 3.3: Corpus-Based Analysis Framework

CHAPTER 4

RESULTS OF THE QUANTITATIVE ANALYSIS

4.0 Introduction

This chapter presents the results of the quantitative analysis, categorised under the following headings:

i) *BE* in the essays written by L1-Malay ESL learners with those written by native speaker learners of English

ii) The correct use of the forms and functions of *BE* by the L1-Malay learners

iii) The incorrect use of the forms and functions of *BE* by the L1-Malay learners

iv) The influence of syntactic environments on the correct and incorrect uses of *BE*

It is important to note that this study employs descriptive statistics for the quantitative analysis administered. It involves measuring the frequency counts of the occurrences of *BE* in forms and functions and presenting these counts into ratios and percentages. The ratio represents the average use of *BE* after the data are normalised to per million words (pmw), while the percentages show the representation of *BE* against the total occurrences of *BE* (finite and non-finite) in each learner corpus. Descriptive statistics is commonly employed in corpus linguistic studies describing patterns of use, for example Hunston (2002) in describing the different verb patterns in English reported the instances of the patterns in ratios. The total frequency of verb with preposition was reported as 134.5 pmw in the British spontaneous spoken English corpus and Guardian newspaper corpus. Similarly, Mohd Don and Sriniwass (2017) also administered

descriptive statistics in the analysis of conjunctive adjuncts in the learner data they analysed. In fact many other studies reviewed in the Chapter 2 employed descriptive statistics in analysing their corpora (e.g. Abdullah & Noor, 2013; Arjan, Abdullah & Roslim, 2013; Hyland & Tse, 2004; Gilquin & Granger, 2015; Leech, 1999; Nessalhauf, 2003), proving that findings from the descriptive statistics can be sufficiently used to determine the patterns of use. For this reason the current study will employ descriptive statistics in its analysis of all the use of *BE* in the learner corpora involved in the study.

4.1 *BE* in the Essays Written by L1-Malay ESL Learners and the Native Learners of English

This section presents the results of the comparative analysis of the use of *BE* between L1-Malay learner sub-corpus and LOCNESS, hence addressing the first research question. It will first present the distributional patterns of *BE* in the native speaker (NS) learner sub-corpora. As already mentioned in Chapter 3, LOCNESS comprises of argumentative essays written by British and American undergraduates. For the purpose of this study, it is divided to American learner sub-corpus and British learner sub-corpus to represent the NS language. The NS learner sub-corpora are analysed separately in order to obtain the overall distribution of *BE* in both the NS learner sub-corpora in the forms of tokens, ratios and percentages. These findings are then used as the basis for the decision either to merge the NS learner sub-corpora to represent a single English variety or to remain as separate sub-corpora, thus, representing two varieties of English.

Past research has attested differences in several grammatical aspects of the two English varieties, for instance in the use of tag questions (Tottie & Hoffmann, 2006), vocatives (Leech, 1999) and verb complementation (Olofsson, 2004). Tottie and Hoffmann (2006) reported that British English has nine (9) times as many tag questions than American

English. Leech (1999) highlighted that American English has higher density of vocatives or 25% more frequent than in British English. British English was also noted to have much greater incidence of kin terms, while American English had much greater incidence of familiarisers. Based on the findings of previous studies there are potential effects of variety (American vs. British variety) in the NS data. In order to determine if there are significant differences in the use of *BE* between the NS learners, this study will first conduct a comparative analysis between the NS learner sub-corpora.

4.1.1 Distribution of *BE* in the American and British Learner Sub-Corpora

This section provides the overall distribution of overt *BE* in LOCNESS in forms of tokens, ratios and percentages of each *BE* form.

The analysis includes all contracted and uncontracted finite *BE* (*am, is, are, was, were , 'm, 's, 're, wasn't, weren't*) and non-finite *BE* forms (*be*, *been*, *being*). To provide a clearer picture of the overall use of both finite and non-finite *BE* forms by the NS learners, the percentage of use is calculated by dividing the number of tokens of each *BE* form by the total tokens of both finite and non-finite *BE* forms. The analysis has excluded forms with zero occurrences. Table 4.1 summarises the tokens, ratios and percentages of *BE* in the British and American sub-corpora.

Table 4.1: Distribution of *BE* in the Bri. and Ame. Learner Sub-Corpora

	Bri.			Ame.		
	Token	**Ratio**	**%**	**Token**	**Ratio**	**%**
am	0	0	0.0	32	214	0.4
‘m	0	0	0.0	0	0	0.0
is	333	16916	**37.2**	2674	17855	**37.3**
‘s	12	610	1.3	160	1068	2.2
isn't	0	0	0.0	0	0	0.0
are	101	5131	**11.3**	1389	9274	**19.4**
‘re	0	0	0.0	0	0	0.0
aren't	0	0	0.0	0	0	0.0
was	47	2388	**5.2**	629	4200	**8.8**
wasn't	0	0	0.0	0	0	0.0
were	26	1321	**2.9**	364	2430	**5.1**
weren't	0	0	0.0	0	0	0.0
be	290	14732	**32.4**	1383	9234	**19.3**
been	50	2540	5.6	317	2117	4.4
being	37	1880	4.1	215	1436	3.0
Total	**896**	**45517**	**100.0**	**7163**	**47828**	**100.0**

As shown in Table 4.1, there are several similarities in the distribution of *BE* across the NS learner sub-corpora. Firstly, both sub-corpora record higher ratios and percentages of finite *BE* compared to non-finite *BE*. The British learner sub-corpus records the overall ratio of 26365pmw (57.9%) of finite *BE* and 19152pmw (42.1%) of non-finite *BE*, while the American learner sub-corpus records the ratio of 35041pmw (73.2%) of finite *BE* and 12787pmw (26.7%) of non-finite *BE*. Secondly, within the finite *BE* category, both British and American sub-corpora record higher percentage of the use of *is* (approximately 37%) than other finite forms. In addition, there is also higher use of the present tense forms than the past tense forms in both NS learner sub-corpora. The combined ratio of all the present *BE* forms in the British learner sub-corpus is 22657pmw (49.85%) compared to 3709pmw (2.9%) of the past forms. In the American learner sub-corpus the present forms record the ratio of 28411pmw (59.3%) compared to 6630pmw (13.9%) of the past forms. Thirdly, both sub-corpora also share similar trend in the use of non-finite *BE*. Infinitive *be* occurs most frequently followed by *been* and *being*. In the British learner sub-corpus infinitive *be* records the ratio of 14732pmw followed by *been* and *being* with the ratios of 2540pmw and 1880pmw respectively,

while the American learner sub-corpus records the ratios of 9234pmw, 2117pmw and 1436pmw of infinitive *be, been* and *being* respectively.

Despite all the similarities highlighted above, there are several differences across the sub-corpora. Firstly, there are differences in the ratios and percentages of the use of infinitive *be*. As presented in Table 4.1, the British learner sub-corpus record the ratio of 14732pmw (32.4%) of infinitive *be* compared to 9234pmw (19.3%) in the American learner sub-corpus. Infinitive *be* is used almost as frequently as *is* in the British learner sub-corpus; 32.4% and 37.2% respectively. Whereas, in the American learner sub-corpus infinitive *be* is the third most frequently used form after *is* and *are*.

The second difference is in the tokens and percentages of *are*. Although *are* is the second most frequently used finite form in both learner sub-corpora, its percentage is higher in the American learner sub-corpus than in the British learner sub-corpus. The American learner sub-corpus has recorded about 19.4% of *are*, while only 11.3% is recorded in the British learner sub-corpus. There is also a larger gap in the percentages of *is* and *are* in the British learner sub-corpus than that in the American learner sub-corpus; about 26% and 18% gaps respectively. The third difference is in the use of past forms *was* and *were*. The American learners use past forms more frequently than the British learners, it records 13.9% combined percentage in the use of past forms compared to only 2.9% recorded in the British learner sub-corpus.

Even though the American and British learners sub-corpora record several similarities in the use of *BE*, there are also key differences which set the two sub-corpora apart. These differences however minute they are, have to be seriously weighed in ensuring that the study provides a true representation of the NS learner language. After taking into consideration all the differences that are evident in the two sub-corpora, a decision

was made to separate the NS learner sub-corpora. Therefore, each will represent a variety of English: the American variety and British variety.

4.1.2 Distribution of *BE* According to Forms in the Learner Sub-Corpora

This section focuses on the comparison of the distribution of *BE* according to forms in L1-Malay learner sub-corpus with that of the British and American learner sub-corpora. The comparison is conducted to discover the similarities and differences in the patterns in the use of *BE* by L1-Malay ESL learners with the NS learners. Table 4.2 below summarises the findings.

Table 4.2: Distribution of *BE* According to Forms in the L1-Malay, Bri. and Ame. Learner Sub-Corpora

	Bri.			Ame.			L1-Malay		
	Token	**%**	**Rank**	**Token**	**%**	**Rank**	**Token**	**%**	**Rank**
am	0	0.0	-	32	0.4	9	40	0.50	9
is	333	37.2	1	2674	37.3	1	4033	48.91	1
‘s	12	1.3	8	160	2.2	8	140	1.70	7
are	101	11.3	3	1389	19.4	2	2170	26.32	2
‘re	0	0.0	-	0	0.0	-	12	0.15	10
was	47	5.2	5	629	8.8	4	344	4.17	4
were	26	2.9	7	364	5.1	5	213	2.58	6
be	290	32.4	2	1383	19.3	3	1028	12.47	3
been	50	5.6	4	317	4.4	6	214	2.60	5
being	37	4.1	6	215	3.0	7	51	0.62	8
Total	**896**	**100.0**		**7 163**	**100.0**		**7 732.0**	**100.0**	

L1-Malay learners’ use of *BE* concentrates mainly on the finite *BE,* in particular *is* and *are*. As shown in Table 4.2 the combined percentage of *is* and *are*, which is at approximately 75.23%, constitutes a major portion of the overall use of *BE*. The non-finite *BE* forms only constitute about 15.62% of the overall use. The difference in percentage between finite and non-finite *BE* uses is about 59.61%.

In contrast, the British learners use finite and non-finite *BE* forms in an almost equal proportion. The combined percentage of *is* and *are* is approximately 48.5%, while the combined percentage of non-finite *BE* forms is at approximately 42.1%. In the

American learner sub-corpus, *is* and *are* record a combined percentage of about 56.7%, whereas the non-finite *BE* forms combined percentage is approximately 26.7%, which results in about 30% difference. The difference in the proportion of finite and non-finite *BE* use across the sub-corpora seem to suggest lack of variation in the use of *BE* among the L1-Malay learners. They tend to be confined to the finite *BE* constructions, unlike their British counterparts, who appear to utilise a more equal proportion of finite and non-finite *BE* forms.

Despite having different proportion of use of the finite and non-finite *BE*, the overall pattern of use of *BE* in the L1-Malay learner data has a very close resemblance to the overall pattern of use of the American learners. As shown in Table 4.2, almost all the *BE* forms are ranked in the same order (except for *being* and *'s*) in both sub-corpora. However, the percentage of each *BE* form differs, for instance in the L1-Malay sub-corpus *is* and *are* each records a percentage of about 48.91% and 26.32% respectively compared to 37.3% and 19.4% in American learner sub-corpus. This finding seems to suggest that the overall use of *BE* of L1-Malay learners is somewhat similar to that the American learners. Nevertheless, they are by no means identical, as highlighted earlier the *BE* forms may appear in the same ranking, but the frequency in their use is clearly different.

Table 4.2 also displays different patterns of the use of *BE* in the British learner sub-corpus compared to the L1-Malay learner sub-corpus. They not only differ in the overall pattern of use but also in the intensity of use. As shown in Table 4.2, the highest ranking forms in the L1-Malay learner data; *is, are* and infinitive *be,* record the percentage of use of 48.91%, 26.23%, and 12.47% respectively, compared to 37.2%, 11.3% and 32.4% recorded in the British learner data.

In sum, the L1-Malay learners overall use of *BE* can be concluded as having an almost similar pattern with that the American learners. This can be clearly seen by the similar ranking of the *BE* forms in both the learner sub-corpora. In contrast, the patterns of use between L1-Malay learners and their British peers are very different, especially in terms of ranking and in the proportion of the use of finite and non-finite *BE* forms.

4.1.3 Distribution of *BE* According to Functions in the Learner Sub-Corpora

In order to further ascertain the similarities and differences in the patterns of the use of *BE* by the L1-Malay learners and the NS learners, the distribution of the *BE* according to the functions it performs was calculated. *BE* was analysed according to their three main functions: copular, auxiliary passive and auxiliary progressive. The calculation includes the four major finite *BE* forms namely; *is, are, was* and *were*. Table 4.3 summarises the findings.

Table 4.3: Distribution of *BE* According to Functions in the L1-Malay, Bri. and Ame. Learner Sub-Corpora

	Ame.			Bri.			L1-Malay		
	Token	**Ratio**	**%**	**Token**	**Ratio**	**%**	**Token**	**Ratio**	**%**
Copular	3893	25994	54.3	414	21031	46.2	5252	26490	63.6
Aux-Passive	799	5335	11.2	76	3861	8.5	1023	5160	12.4
Aux-Progressive	364	2430	5.1	17	864	1.9	497	2507	6

As shown in Table 4.3, L1-Malay learners and the native speaker learners exhibit similar trend in the overall use of *BE*. Across the sub-corpora copula *BE* constructions occur the most, followed by auxiliary *BE* in passive voice and auxiliary *BE* in progressive aspect. Nevertheless, L1-Malay learners appear to record the highest percentage of copula *BE* constructions compared to the NS learners. Approximately 63.6% of the total use of *BE* constitute copular constructions compared to only 46.2% of the same constructions in the British learner sub-corpus and 54.3% in the American learner sub-corpus. This seems to suggest a heavier reliance on copula *BE* constructions by the L1-Malay learners when compared to the native speaker learners.

As for auxiliary constructions, there appear to be similarities in the percentages of use of auxiliary *BE* in the passive and progressive constructions by the L1-Malay and American learners. As can be seen in Table 4.3, the percentage of passive constructions by L1-Malay learners is at 12.4%, which is close to the 11.2% recorded in the American learner sub-corpus. The same tendency is also exhibited in the construction of progressives, the L1-Malay learner sub-corpus records 6% use of auxiliary *BE* in marking progressive aspect, while the American learner sub-corpus records a slightly lower percentage of 5.1% for the same function. These figures suggest that the overall use of auxiliary *BE* by the L1-Malay learner is closer to that the American learners than the British learners. It is also interesting to note that, the British learners record the lowest percentages of use both copula *BE* and auxiliary *BE*; approximately 46.2% and 10.4% respectively. This finding suggests that the British learners do not rely as heavily on the use of copula *BE* and auxiliary *BE* constructions as would the L1-Malay learners.

In general, the L1-Malay learners are found to use more copular constructions in their writings compared to the native speaker learners. Nevertheless, the overall pattern of the use of auxiliary *BE* in the progressive aspect and passive voice is very similar to the overall pattern recorded in the American learner sub-corpus.

4.1.4 Summary of the Patterns of the Use of *BE* in the Learner Sub-Corpora

The patterns of the use of *BE* in the learner sub-corpora can be summarised as:

1. All the learner sub-corpora share four common traits; they have (a) higher occurrences of finite than non-finite forms, (b) higher occurrences of present forms than past forms, (c) higher occurrences of singular forms than plural forms and (d) higher occurrences of copula *BE* than auxiliary *BE*.

2. L1-Malay learners' use of *BE* concentrates heavily on the finite forms *is* and *are,* while the native speaker learners tend to use finite and non-finite forms more equally.
3. In terms of the major functions, L1-Malay learners appear to use more copular constructions compared to the native speaker learners.
4. The overall pattern of the use of *BE* by the L1-Malay learners is closer to the overall pattern of the American learners than the British learners.

4.2 Grammatical Use of *BE* in the L1-Malay Learner Sub-Corpus

This section presents the findings of the quantitative analysis of the grammatical use of *BE* in the L1-Malay learner sub-corpus. The findings are sequenced according to the research questions.

4.2.1 Distribution of Grammatical use of *BE* According to Forms and Functions

In this section the focus is on the distribution of the grammatical use of each form and function of *BE.*

4.2.1.1 Distribution of Grammatical Use of *BE* According to Forms

In order to gauge the overall pattern of the use of *BE,* each form of *BE* was analysed for its grammaticality. Table 4.4 summarises the findings. The ratio represents the average number occurrences per a million English words (pmw). This method of calculating the average of the tokens allows comparison of the distribution of *BE* within and between corpora to be carried out.

Table 4.4: Distribution of Grammatical *BE* According Forms in L1-Malay Data

Finite *BE*				Non-Finite *BE*			
***be* Forms**	**Token**	**Ratio**	**%**	***be* Forms**	**Token**	**Ratio**	**%**
am	40	201.75	0.48	be	1028	5185.06	12.45
'm	0	0.00	0	been	214	1079.38	2.59
is	4039	20372.03	48.92	being	51	257.24	0.62
's	137	696.05	1.67				
isn't	0	0.00	0				
are	2173	10960.24	26.32				
're	12	60.53	0.15				
aren't	0	0.00	0				
was	345	1740.12	4.18				
wasn't	0	0.00	0				
were	215	1084.42	2.60				
weren't	1	5.04	0.01				
Total	**6962**	**35120.19**	**84.33**		**1293**	**6521.67**	**15.66**

As shown in Table 4.4, finite *BE* occurs more frequently than non-finite *BE*. Approximately 84.33% of the total occurrences of *BE* constitute the finite forms, while only about 15.66% constitute the non-finite forms. Within the finite *BE* forms, there is a vast difference in the occurrences of present forms than the past forms. The present *BE* forms *is* and *are* record the two highest ratios; *is* with 20372.03pmw, followed by *are* with 10960.24pmw, while the past forms *was* and *were* record the ratios of 1740.12pmw and 1084.42pmw respectively. In addition, there also exists a gap of 22.6% between singular form *is* and plural *are*, suggesting higher use of singular subjects in the L1-Malay learner data. The table also displays very limited use of the present form *am*, which has recorded the ratio of only 201.75pmw. This figure is expected in learners' argumentative writing, where the use of personal pronoun *I* is usually restricted to stance making in phrases such as *I agree/disagree...*, *I think that...* or *I believe that .... .*

As mentioned earlier, non-finite forms are used less frequently than the finite forms; infinitive *be* records the ratio of 5185.06pmw, while *been* and *being* record the ratios of 1079.38pmw and 257.24pmw respectively. Table 4.4 also displays a very minimum use of contracted *BE* forms; 696.05pmw of *'s*, 60.53pmw of *'re* and 5.04pmw of *weren't*.

Other contractions, namely *'m, isn't aren't and wasn't,* are not found in the data. This finding is expected as the entire corpus contains only of academic essays and contracted forms are less favoured in academic writing (Biber et al., 1999). As can be seen from Table 4.4, three forms; *is*, *are* and infinitive *be*, have recorded comparatively higher percentages of use and they are also ranked as the three most used forms; 48.92% of *is*, followed by 26.32 % of *are* and 12.45% of infinitive *be.*

The overall distribution of *BE* forms in the L1-Malay learner sub-corpus can be summarised as:

1. Finite *BE* forms occur more frequently than non-finite *BE* forms.
2. The present forms are used more frequently than the past forms.
3. The singular forms are used more frequently than the plural forms.
4. Three most frequently used *BE* forms are; *is*, *are* and infinitive *be*.

4.2.1.2 Distribution of Grammatical Use of *BE* According to Functions

The distribution of *BE* here is discussed in terms of the functions it performs, namely as copular, auxiliary, negative operator, interrogative operator and in existential *there* and *it*-clefts. Table 4.5 summarises the findings.

Table 4.5: Distribution of Finite *BE* According to Functions in L1-Malay Data

Functions	Token	Ratio	%
Copular	4287	21 622.90	61.58
Auxiliary	1387	6 995.80	19.92
Negation operator	424	2 138.58	6.09
Interrogative operator	81	408.55	1.16
Existential *there*	544	2 743.84	7.81
It-cleft	239	1 205.48	3.43

As shown in Table 4.5, more than half or approximately 61.58% of *BE* in the L1-Malay learner sub-corpus function as copular. The high percentage of copula *BE* in the corpus is expected as copula *BE*, according to Biber et al. (1999) is more common in the

written register and in academic prose. Copula *BE* can also occur in a wide range of complements; noun phrase, adjective phrase, prepositional phrase and complement clause (Biber et al., 1999), which could contribute to its relatively higher occurrences in the learner sub-corpus .

Approximately 19.92% of finite *BE* in the sub-corpus function as auxiliary to mark progressive aspect (*BE + Ving*) or in the formation of passives (*BE + Ven*). The lower percentage of auxiliary *BE* shows that progressives and passives are used less frequently by the Malay learners compared to copula *BE* constructions. Table 4.5 also displays very limited use of *BE* performing other functions besides copula and auxiliary. Negative and interrogative operators record 6.09% and 1.16% of use respectively, while existential *there* and *it*-cleft record low percentages of approximately 7.81% and 3.43% respectively. To gain better insights on the distributional patterns of finite *BE*, the tokens and percentages are also presented according to forms and functions. Table 4.6 summarises the distribution of the major finite *BE* forms (*is, are, was, were*) according to their functions.

Table 4.6: Distribution of Finite *BE* According to Forms and Functions in L1-Malay Data

Form/	**is**		**are**		**was**		**were**	
Functions	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
Copula	2800	40.20	1096	15.70	156	2.20	81	1.16
Auxiliary	472	6.80	669	9.60	149	2.10	97	1.39
Negative operator	249	3.60	144	2.10	10	0.10	10	0.14
Interrogative operator	70	1.00	11	0.20	0	0.00	0	0.00
Existential there	234	3.40	253	3.60	24	0.30	27	0.39
It-cleft	214	3.10	0	0.00	6	0.10	0	0.00
Total	4039	58.10	2173	31.20	345	4.80	215	3.08

Table 4.6 displays that 58.10% of the finite *BE* used by the L1-Malay learners constitute *is* and most of them function as copula in a main verb position (40.20%) and only about 6.8% function as auxiliary, 3.6% as negative operators, 1% as interrogative operators and 3.4% in existential *there* and 3.1% in *it*-cleft constructions. There also exists a huge

gap between the use of *is* as copula and auxiliary; the former exceeds the latter by approximately 33.4%. It is clear from this figure that the majority of *is* in this study functions as copula and the higher occurrences of *is* also indicate more frequent use of singular subjects in the copula *BE* constructions.

The second most used finite form *are* records 15.7% of usage as copula, compared to about 9.6% as auxiliary, 2.10% as negative operators and 3.6% in the existential *there* constructions. There exists comparatively smaller gap between the percentage of copula and auxiliary *are* (6%), suggesting that the use of *are* is not restricted to mainly copular constructions. In addition, compared to other forms *are* also recorded the highest percentage of use as auxiliary; 9.6% compared to 6.8% recorded for *is*, 2.10% for *was* and 1.39% for *were*.

The distributional patterns of finite *BE* according to functions in L1-Malay learner sub-corpus can be summarised as:

1. Finite *BE* functions primarily as a copular in a main verb position and secondly as an auxiliary.
2. The form *is* records the highest occurrences and functions mainly as a copular.
3. The form *are* is the second most used form and it has recorded the highest percentage of use as an auxiliary.
4. The present forms (*is, are*) exceed the past forms (*was, were*) for all the major functions (copular and auxiliary).

4.2.2 Patterns of Grammatical Use of Finite *BE*

This section reports the use of all the major functions of finite *BE*, which will then unveil the patterns of use of each function.

4.2.2.1 Patterns of the Use of Copula *BE*

In order to reveal the patterns of use of copula *BE*, this section explores in-depth the occurrences of copula *BE* in relation to the constituents occurring before and after *BE*. The constituents refer specifically to the type of subjects and subject predicates used in the copula *BE* constructions.

4.2.2.1.1 Type of Subjects

The types of subjects are categorised into lexical noun (NP), personal pronoun (PPN), definite pronoun (DPN), indefinite pronoun (IPN) and *wh*-pronoun (QPN). Lexical nouns refer to any form of nouns used by the learners as the subject of the copula *BE* sentences, they include single-word noun such as *money* or a noun phrase such as *Many higher institutions in Malaysia*. Personal pronouns include the pronouns *I, we, she, he, it, they* and *you*, definite pronouns include *that, this, these* and *those*, indefinite pronouns include any compound pronouns such as *somebody, everything, anyone*, quantifiers such as *some, all*, *any* or the pronoun *one* and finally *wh*-pronouns include any *wh*-word that is used as a subject in a copula-*BE* sentence. In order to obtain a clear picture of the distribution of the different subjects in the copula *BE* constructions, the analysis is conducted for every finite *BE* form. However, *am* was excluded from the analysis as it only takes the pronoun *I* as its subject. Table 4.7 presents the findings.

Table 4.7: Distribution of Type of Subjects in Copula *BE* Constructions in L1-Malay Data

	Noun (NP)		Personal Pronoun (PPN)		Definite Pronoun (DPN)		Indefinite Pronoun (IPN)		Wh-Pronoun (QPN)	
	Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
is	1811	42.24	564	13.16	274	6.39	35	0.82	116	2.71
‘s	6	0.14	62	1.45	35	0.82	2	0.07	8	0.19
are	770	17.96	258	6.02	48	1.12	10	0.23	10	0.23
‘re	1	0.02	11	0.26	0	0.00	0	0.00	0	0.00
was	87	2.03	52	1.21	9	0.21	2	0.05	6	0.14
were	52	1.21	26	0.61	2	0.05	0	0.00	1	0.02
Total	**2727**	**63.61**	**973**	**22.70**	**368**	**8.58**	**50**	**1.17**	**141**	**3.29**

As shown in Table 4.7, the use of lexical nouns (NP) dominates the copula *BE* constructions in this study. The overall use of NP for all *BE* forms is recorded at 63.61%. The second most used subject is personal pronouns (PPN) with 22.7% of occurrences. Definite pronouns (DPN), indefinite pronouns (IPN) and *wh*-pronouns (QPN) record relatively low percentages of use of 8.58%, 3.29% and 1.9% respectively. All the *BE* forms, except for *'s* and *'re*, tend to favour NP subjects. The contracted forms *'s* and *'re* record higher occurrences with PPN subjects and this is highly expected as *BE* forms are usually contracted with pronouns for instance *he's, she's* and *they're.*

In general, the results from the analysis of the types of subjects for copula *BE* constructions reveal that *NP + BE* structure occurs most frequently in the L1-Malay learner data followed by *PPN + BE* structure.

4.2.2.1.2 Subject Predicates

Copula *BE* occurs with a wide range of complements. These complements can either be phrases, namely noun phrase, adjective phrase and prepositional phrase, or complement clauses, namely infinitive-*to* clause, *that*-clause and *wh*-clause. This study found occurrences of all these predicates in the L1-Malay learner data as summarised in Table 4.8 below:

Table 4.8: Distribution of Subject Predicates in the Copula *BE* Constructions in L1-Malay Data

	Noun Phrase (NP)		Adjective Phrase (AP)		Prepositional Phrase (PP)		Infinitive Clause (InfC)		That Clause (ThtC)		Wh-Clause (WhC)	
	Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
is	1373	32.03	756	17.63	210	4.90	104	2.43	63	1.47	74	1.73
's	42	0.98	44	1.03	10	0.23	0	0.00	1	0.02	16	0.37
are	513	11.97	424	9.89	117	2.73	15	0.35	0	0.00	3	0.07
're	2	0.05	3	0.07	7	0.16	0	0.00	0	0.00	0	0.00
was	77	1.80	56	1.31	14	0.33	4	0.09	1	0.02	2	0.05
were	34	0.79	31	0.72	8	0.19	6	0.14	0	0.00	0	0.00
Total	**2041**	**47.61**	**1314**	**30.65**	**366**	**8.54**	**129**	**3.01**	**65**	**1.52**	**95**	**2.22**

As shown in Table 4.8, copula *BE* constructions in the L1-Malay learner data are most frequently complemented by NP (47.61%), and followed by AP (30.65%), while PP records a low percentage of only 8.54%. Clauses in general are very marginally used as complements; infinitive *to*-clause with 3.01%, *wh*-clause with 2.22% and *that*-clause with 1.52%. These findings are consistent with the findings of Lee and Huang (2004), who also recorded higher frequency of correct use *BE*-noun and *BE*-adjective constructions in the L1-Chinese learner data they investigated.

The findings reveal firstly, that the majority of the copula *BE* constructions in the L1-Malay data tend to be complemented by phrases (86.8%) rather than clauses (6.75%). Secondly, a significant number of copula *BE* constructions in the L1-Malay learner data are constructed with NP and AP predicates with a combined percentage of 78.3%. Hence, it can be concluded that *BE + NP* and *BE + AP* structures are most common in the L1-Malay learner sub-corpus.

4.2.2.1.3 Summary of the Patterns of Use of Copula BE

The patterns copula *BE* constructions in the L1-Malay learner sub-corpus can be summarised as:

1. Copula *BE* occurs primarily with NP subjects followed by PPN subjects with NP subjects occurring more frequently than PPN subjects (*NP + BE >PPN + BE*).
2. NP and AP predicates are used most frequently as complements of copula *BE* constructions with NP predicates occurring more frequently than AP predicates (*BE + NP > BE + AP*).
3. In general the copula *BE* constructions in the L1-Malay learner sub-corpus are likely to occur as follows:
   a) *NP + BE + NP*
   b) *PPN + BE + NP*

c) *NP + BE + AP*

d) PPN + BE + AP

4.2.2.2 Patterns of the Use of Auxiliary *BE*

This section discusses in detail the patterns of the grammatical use of auxiliary *BE* in the L1-Malay learner sub-corpus. It begins by presenting the distribution of auxiliary progressive and auxiliary passive in the L1-Malay learner data, followed by the distribution of the constituents surrounding auxiliary *BE*, which include the type of subjects and the class of post-*BE* verbs.

4.2.2.2.1 Distribution of Auxiliary *BE* in Progressive and Passive Constructions

Auxiliary *BE* has two major grammatical functions that are to mark progressive aspect, which is realised as *BE + Ving* (*They <u>are seeking</u> better job opportunities*) and to form passives in *BE + Ved/en* construction (*The proposal <u>was approved</u> by the Head of the Department*). This section presents the distribution of auxiliary *BE* that are used as progressive aspect marker and in the formation of passives. The findings are presented in Table 4.9.

Table 4.9: Distribution of Auxiliary *BE* in Progressive and Passive Constructions in L1-Malay Data

Form	Aux-Progressive *BE + Ving*		Aux-Passive *BE + Ved*	
	Token	**%**	**Token**	**%**
is	125	9.01	347	25.02
are	287	20.69	382	27.54
was	26	1.87	123	8.87
were	13	0.94	84	6.06
Total	**451**	**32.52**	**936**	**67.48**

As shown in Table 4.9, auxiliary *BE* is used more frequently in the construction of passives (67.48%) than as progressive aspect markers (32.52%). This finding shows more frequent use of passives in the L1-Malay learners' argumentative essays as opposed to progressives. This empirical evidence is consistent with the findings of

Biber et al. (1999), who also reported more common use of passive voice in academic prose compared to progressive aspect.

4.2.2.2.1 Type of Subjects

This section reports the findings of the type of subjects occurring with the auxiliary *BE* in the L1-Malay learner data. They include nouns, personal pronouns, definite pronouns, indefinite pronouns and *wh*-pronouns. Table 4.10 summarises the findings.

Table 4.10: Distribution of Type of Subjects in Auxiliary *BE* Constructions in L1-Malay Data

		Noun (NP)		Personal Pronoun (PPN)		Definite Pronoun (DPN)		Indefinite Pronoun (IPN)		Wh-Pronoun (Wh-P)	
		Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
Aux-Progressive	is	85	18.85	15	3.33	4	0.89	8	1.77	13	2.88
	are	137	30.38	128	28.38	0	0.00	0	0.00	22	4.88
	was	9	2.00	13	2.88	0	0.00	0	0.00	4	0.89
	were	6	1.33	7	1.55	0	0.00	0	0.00	0	0.00
	Total	**237**	**52.55**	**163**	**36.14**	**4**	**0.89**	**8**	**1.77**	**39**	**8.65**
Aux-Passive	is	244	26.07	68	7.26	11	1.18	9	0.96	15	1.60
	are	285	30.45	84	8.97	9	0.96	2	0.21	2	0.21
	was	88	9.40	33	3.53	0	0.00	0	0.00	2	0.21
	were	63	6.73	18	1.92	2	0.21	0	0.00	1	0.11
	Total	**680**	**72.65**	**203**	**21.69**	**22**	**2.35**	**11**	**1.18**	**20**	**2.14**

Table 4.10 shows that as progressive aspect markers, approximately 52.55% of auxiliary *BE* occurs with lexical nouns (NP) and about 36.14% with personal pronouns (PPN), while *wh*-pronouns (QPN), definite (DPN) and indefinite pronouns (IPN) record very low percentage of 8.65%, 0.89% and 1.77% respectively. As for *BE* in the passive constructions, 72.65% occur with NP, 21.69% with PPN and 2.35%, 1.18% and 2.14% occur with DPN, IPN and QPN respectively.

NP subjects are used most frequently with both passive and progressive constructions, but they tend to occur more frequently in the passive constructions than in the progressive constructions. PPN tend to occur slightly more frequent in the progressive aspect than in the passive voice. In general, it is found that NP and PPN are the most

preferred types of subjects in both progressive and passive constructions with NP subjects occurring more frequently than PPN subjects (NP>PPN).

4.2.2.2.3 Class of Post-BE Lexical Verbs

Class of post-*BE* lexical verbs is an integral part of the analysis of progressive and passive constructions in the L1-Malay ESL learner data. The lexical verbs are categorised into transitive (Vt), unergative (Uer) and unaccusative (Uac) verbs. The findings are summarised in Table 4.11 below:

Table 4.11: Distribution of Post-*BE* Verbs According to Class in Auxiliary *BE* Constructions in L1-Malay Data

		Transitive (Vt)		Unergative (Uer)		Unaccusative (Uac)	
		Token	**%**	**Token**	**%**	**Token**	**%**
Aux-Progressive	is	73	16.19	31	6.87	21	4.66
	are	173	38.36	99	21.95	15	3.33
	was	13	2.88	13	2.88	0	0.00
	were	9	2.00	4	0.89	0	0.00
	Total	**268**	**59.42**	**147**	**32.59**	**36**	**7.98**
Aux-Passive	is	347	37.07	0	0.00	0	0.00
	are	382	40.81	0	0.00	0	0.00
	was	123	13.14	0	0.00	0	0.00
	were	84	8.97	0	0.00	0	0.00
	Total	**936**	**100**	**0**	**0.00**	**0**	**0.00**

Auxiliary *BE* in progressive constructions tends to be proceeded most often by transitive (Vt) and unergative (Uer) verbs. However, Vt verbs occur more frequently (59.42%) than Uer verbs (32.59%), while Uac verbs record the lowest percentage of use of only 7.98%.

As for the passive constructions, only Vt verbs (100%) are used in the formation of passives. This is mainly because only transitive verbs can be passivised. Unergative and unaccusative verbs only take one argument. The sole argument of an unergative verb maps onto the subject position and it does not require any object as in *Mary danced* (Park & Lakshmanan, 2007, p. 329). Unergative verbs, therefore, could not be passivised. As for unaccusative verbs, the sole argument of an unaccusative verb, which

is the theme, maps onto the subject position as in *The snow melted* (Park & Lakshmanan, 2007, p. 329). Unaccusative construction does not have a causal agent, making it impossible to passivise. For these reasons both unergative and unaccusative verbs are not used in the constructions of passives.

In general, progressives in the L1-Malay learner data occur mainly with Vt and Uer verbs, while the passive constructions occur exclusively with Vt verbs.

4.2.2.2.4 Summary of the Patterns of Use of Auxiliary BE

The patterns of the use of auxiliary *BE* in the L1-Malay learner sub-corpus can be summarised as:

1. Auxiliary *BE* is used more frequently in the formation of passives than progressives (AuxPas>AuxProg).
2. In the formation of progressives, NP and PPN subjects are used most frequently, with NP subjects occurring more frequently than PPN subjects. Vt and Uer verbs are mainly used in the progressives, with Vt verbs occurring more frequently than Uer verbs. The auxiliary *BE* progressive constructions in the L1-Malay learner sub-corpus are likely to occur as follows:
    a) *NP + BE + Vt-ing,*

       NP + BE + Uer-ing,

       PPN + BE + Vt-ing ,

       PPN + BE + Uer-ing,
3. In the formation of passives, only NP and PPN are used as subjects, with NP subjects used more frequently than PPN subjects. As for the class of post-*BE* verbs only Vt verbs are used. The passives in this study are likely to occur as follows:

a) *NP + BE + Vt-ed*

PPN + BE + Vt-ed

4.2.2.3 Patterns of the Use of *BE* as Negation Operator

Another function of *BE* is as an operator for clause negation. The verb can either perform the task of negation operator in copular clauses (*You**'re not** pretty*) (Biber et al, 2002, p. 238) or negation operator in auxiliary clauses (*They **are not forgotten**/I **am not looking** for an employee of yours*) (Biber et al., 2002, p.162). In this study the auxiliary operator is further divided to progressive and passive auxiliary operators. This section presents the tokens and percentages of both copula *BE* and auxiliary *BE* used as negation operators in the L1-Malay ESL learner data. Table 4.12 summarises the findings.

Table 4.12: Distribution of Copula and Auxiliary *BE* as Negation Operators in L1-Malay Data

Functions	Copular		Auxiliary Progressive		Auxiliary Passive	
	Token	%	Token	%	Token	%
am not	11	2.57	0	0.00	0	0.00
is not	229	53.50	8	1.87	12	2.80
are not	103	24.07	16	3.74	25	5.84
was not	6	1.40	0	0.00	4	0.93
were not	10	2.34	0	0.00	4	0.93
Total	**359**	**83.88**	**24**	**5.61**	**45**	**10.51**

In general, there is a very low percentage of *BE* used as negation operators in the sub-corpus. The overall percentage is recorded at about 6% (refer to Table 4.5), the majority of which function as negation operators in copular constructions (83.88%), while a small percentage is used to negate auxiliary constructions (16.12%). The low overall percentage of *BE* used as negation operators in the L1-Malay ESL learner data indicates a very limited use of negations by the L1-Malay learners.

4.2.2.3.1 Pre-BE and Post-BE Constituents of BE as Negation Operators

This section presents the findings of the constituents before and after negative *BE*. They include the different type of subjects, subject predicates (copula *BE*) and class of post-*BE* lexical verbs (auxiliary *BE*).

4.2.2.3.2 Type of Subjects

As shown in Table 4.13, NP and PPN are used most frequently as the subjects of negative copula *BE* and auxiliary *BE* constructions. Negative copula *BE* constructions recorded 60.06% occurrences of NP subjects, which almost double the percentage of PPN subjects (29.31%). DPN, IPN and QPN subjects record 6.9% 0.86% and 2.87% of use respectively. As for negative progressive constructions, NP subjects occur slightly more frequent than PPN subjects, 50% and 45.83% respectively. As for the negative passive constructions, NP subjects record about 55.56% of use compared to 42.22% recorded for PPN subjects.

Table 4.13: Distribution of Type of Subjects in Negative Constructions in L1-Malay Data

		Noun (NP)		Personal Pronoun (PPN)		Definite Pronoun (DPN)		Indefinite Pronoun (IPN)		Wh-Pronoun (QPN)	
		Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
Copular	is not	143	41.09	60	17.24	23	6.61	2	0.57	1	0.29
	are not	55	15.80	38	10.92	0	0.00	1	0.29	9	2.59
	was not	3	0.86	2	0.57	1	0.29	0	0.00	0	0.00
	were not	8	2.30	2	0.57	0	0.00	0	0.00	0	0.00
	Total	**209**	**60.06**	**102**	**29.31**	**24**	**6.90**	**3**	**0.86**	**10**	**2.87**
Aux-Progressive	is not	6	25.00	2	8.33	0	0.00	0	0.00	0	0.00
	are not	6	25.00	9	37.50	0	0.00	0	0.00	1	4.17
	was not	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	were not	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Total	**12**	**50.00**	**11**	**45.83**	**0**	**0.00**	**0**	**0.00**	**1**	**4.17**
Aux-Passive	is not	9	20.00	2	4.44	1	2.22	0	0.00	0	0.00
	are not	10	22.22	15	33.33	0	0.00	0	0.00	0	0.00
	was not	4	8.89	0	0.00	0	0.00	0	0.00	0	0.00
	were not	2	4.44	2	4.44	0	0.00	0	0.00	0	0.00
	Total	**25**	**55.56**	**19**	**42.22**	**1**	**2.22**	**0**	**0.00**	**0**	**0.00**

In general, NP is used most frequently as the subjects for negative copula and auxiliary *BE* constructions followed by PPN subjects.

4.2.2.3.3 Subject Predicates (Copula BE)

This section presents the results of the types of predicates complementing negative copula *BE*. The complements include noun phrase (NP), adjective phrase (AP), prepositional phrase (PP), infinitive clause (InfC), *wh*-clause (WhC) and *that*-clause (ThtC). The findings are presented in Table 4.14 below:

Table 4.14: Distribution of Subject Predicates in Negative Constructions in L1-Malay Data

	Noun Phrase (NP)		Adjective Phrase (AP)		Prepositional Phrase (PP)		Infinitive Clause (InfC)		Wh/That Clause (Wh-C/ThtC)	
	Token	%	Token	%	Token	%	Token	%	Token	%
is not	115	32.12	98	27.37	12	3.35	2	0.56	2	0.56
are not	18	5.03	65	18.16	14	3.91	4	1.12	2	0.56
was not	3	0.84	2	0.56	1	0.28	0	0.00	0	0.00
were not	4	1.12	6	1.68	0	0.00	0	0.00	0	0.00
Total	**140**	**39.11**	**171**	**47.77**	**27**	**7.54**	**6**	**1.68**	**4**	**1.12**

AP predicates are used most frequently to complement negative copula *BE*. As can be seen in Table 4.14, 47.77% of the constructions are complemented by AP predicates, while 39.11% are complemented by NP predicates. Prepositional phrase, infinitive *to*-clause and *wh/that*-clause, however, record very low percentages of occurrences; 7.55%, 1.68% and 1.12% respectively. In general, negative copula *BE* in the L1-Malay learner sub-corpus takes mostly AP and NP predicates as complements, with AP predicates occurring more frequently than NP predicates.

4.2.2.3.4 Class of Post-BE Verbs (Auxiliary BE)

Table 4.15 displays the results for the analysis of post-*BE* lexical verbs that occur with the negative progressive and passive constructions.

Table 4.15: Distribution of Class of Post-*BE* Verbs in Negative Constructions in L1-Malay Data

		Transitive (Vt)		Unergative (Uer)		Unaccusative (Uac)	
		Token	**%**	**Token**	**%**	**Token**	**%**
Aux-Progressive	is not	5	20.83	2	8.33	0	0.00
	are not	11	45.83	6	25.00	0	0.00
	was not	0	0.00	0	0.00	0	0.00
	were not	0	0.00	0	0.00	0	0.00
	Total	**16**	**66.67**	**8**	**33.33**	**0**	**0.00**
Aux-Passive	is not	12	26.67	0	0.00	0	0.00
	are not	25	55.56	0	0.00	0	0.00
	was not	4	8.89	0	0.00	0	0.00
	were not	4	8.89	0	0.00	0	0.00
	Total	**45**	**100.00**	**0**	**0.00**	**0**	**0.00**

As presented in Table 4.15, *BE* negating progressive aspect is found to occur only with Vt and Uer verbs; approximately 66.67% and 33.33% respectively, while *BE* in the passive constructions as anticipated only occurs with Vt verbs.

4.2.2.3.5 Summary of the Patterns of Use of BE as Negation Operator

The patterns of the use of *BE* as negation operators in the L1-Malay learner sub-corpus can be summarised as:

1. There are higher constructions of negative copula *BE* than negative auxiliary *BE* in the sub-corpus (NegCop>NegAux).
2. Negative copula *BE* constructions occur mostly after NP and PPN subjects with NP subjects occurring more frequently than PPN subjects. Negative copula *BE* constructions are complemented mostly by AP and NP predicates. The use of AP predicates slightly exceeds the use of NP predicates. The copula *BE* negative constructions are likely to occur as follows:

   a) NP + NegCop BE + AP

   b) PPN + NegCop BE + AP

   c) NP + NegCop BE + NP

   d) PPN + NegCop BE + NP

3. Auxiliary *BE* in negative progressive is preceded mostly by NP and PPN subjects and followed by Vt and Uner verbs, with Vt verbs occurring approximately twice more frequently than Uer verbs. In general, the negative progressives are likely to occur as follows:

    a) *NP + AuxNeg BE + Vt-ing,*

    b) *PPN +AuxNeg BE + Vt-ing,*

    c) *NP + AuxNeg BE + Uner-ing*

    d) *PPN +AuxNeg BE + Uner-ing*

4. Auxiliary *BE* in negative passive constructions occur with NP and PPN subjects and followed only by Vt verbs. They are likely to occur as the followings:

    a) *NP + AuxNeg BE + Vt-ed*

    b) *PPN + AuxNeg BE + Vt-ed*

4.2.2.4 Patterns of the Use of *BE* as Interrogative Operator

BE is an important element in the formation of interrogative clauses. The analysis includes only the use of *BE* in the formation of *yes/no* questions, whereby *BE* functions as an operator and placed in front of the subject NP in the subject-operator inversion (*BE + Sub + Comp?*) as in '*Is that lovely?'* (Biber et al., 2002, p. 251). *BE* can be used as the interrogative operators of copular clauses as in (a) and auxiliary clauses as in (b) and (c):

a) ***Is*** *he happy?* Copular

b) ***Are*** *they coming soon*? Auxiliary Progressive

c) ***Is*** *the ring made of gold*? Auxiliary Passive

Table 4.16 summarises the distribution of *BE* as interrogative operators in the copular and auxiliary constructions.

Table 4.16: Distribution of *BE* as Interrogative Operators in L1-Malay Data

	Copular		Auxiliary	
	Token	**%**	**Token**	**%**
is	66	81.48	4	4.94
are	7	8.64	4	4.94
was	0	0.00	0	0.00
were	0	0.00	0	0.00
Total	**73**	**90.12**	**8**	**9.88**

The overall percentage of *BE* that functions as interrogative operators in the L1-Malay learner sub-corpus is only about 1.16% (refer to Table 4.5). Table 4.16 shows that *BE* in the L1-Malay learner data is used more frequently as interrogative operators in the copular constructions (90.12%) than in auxiliary constructions (9.88%). It is also found that only the present forms (*is, are*) are used in the construction of interrogatives.

4.2.2.4.1 Pre-BE and Post-BE Constituents of BE as Interrogative Operator

This section presents the findings for the pre-*BE* and post-*BE* constituents in the interrogative constructions. They include the type of subjects, subject predicates (copula *BE*) and class of post-*BE* lexical verbs (auxiliary *BE*).

4.2.2.4.2 Type of Subjects

Table 4.17 below summarises the findings of the types of subjects used in interrogative constructions in the L1-Malay learner sub-corpus.

Table 4.17: Distribution of the Type of Subjects in Interrogative Constructions in L1-Malay Data

		Noun (NP)		Personal Pronoun (PPN)		Definite Pronoun (DPN)	
		Token	**%**	**Token**	**%**	**Token**	**%**
copula	is	19	23.46	28	34.57	19	23.46
	are	1	1.23	6	7.41	0	0.00
aux-passive	is	2	2.47	0	0.00	0	0.00
	are	1	1.23	1	1.23	0	0.00
aux-prog	is	2	2.47	0	0.00	0	0.00
	are	0	0.00	1	1.23	1	1.23
	Total	**25**	**30.86**	**36**	**44.44**	**20**	**24.69**

As shown in Table 4.17, only NP, PPN and DPN occur with *BE* in the interrogatives. PPN subjects record the most frequent occurrences, followed by NP and DPN with the

percentages of 44.44%, 30.86% and 24.69% respectively. In general interrogatives in the L1-Malay learner data take mostly PPN and NP subjects.

4.2.2.4.3 Subject Predicates

The analysis of the subject predicates was only administered for interrogative operator in copula *BE* constructions. Table 4.18 presents the findings.

Table 4.18: Distribution of Subject Predicates in Interrogative Constructions in L1-Malay Data

	Adjective Phrase (AP)		**Noun Phrase (NP)**		***Wh*-clause (Wh-C)**		**Adverbial Clause (AdvC)**	
	Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
is	30	41.10	22	30.14	8	10.96	6	8.22
are	4	5.48	3	4.11	0	0.00	0	0.00
Total	**34**	**46.58**	**25**	**34.25**	**8**	**10.96**	**6**	**8.22**

Table 4.18 displays that only four subject predicates are used in the copula *BE* interrogatives, they include adjective phrase (AP), noun phrase (NP), *wh*-clause (WhC) and adverbial clause (AdvC). AP and NP predicates, which recorded the percentages of 46.58% and 34.25% respectively, are used most frequently as complements, while AdvC and WhC record very low occurrences of 10.96% and 8.22% respectively. In general, the interrogative constructions in the L1-Malay learner sub-corpus occur mostly with AP and NP predicates, with AP predicates occurring slightly more frequent than NP predicates.

4.2.2.4.4 Class of Post-BE Lexical Verbs

Since there are only eight (8) tokens of *BE* used as interrogative operators in the auxiliary constructions, the analysis for post-*BE* lexical verbs was not administered.

4.2.2.4.5 Summary of the Patterns of the Use of BE as Interrogative Operators

The patterns of the use of *BE* as interrogative operators in the L1-Malay learner sub-corpus can be summarised as follows:

1. *BE* is used most frequently as interrogative operators in the copular constructions.

2. The interrogatives are most frequently preceded by PPN subjects, followed by NP subjects. AP and NP predicates are most frequently used as complements of copula *BE* interrogatives. They are likely to occur as follows:

    a. *BE + PPN + AP?*

    b. *BE + NP + AP?*

    c. *BE + PPN + NP?*

    d. *BE + NP + NP?*

4.2.2.5 Patterns of the Use of *BE* in Existential *there*

This section discusses the function of *BE* in a unique existential *there* structure. Existential *there* commonly takes the structure *there + BE + NP + adverbial* (*There are around 6 000 accidents in the kitchen of Northern Ireland homes every year*) in which *BE* is usually a main verb (Biber, et al., 2002). Nevertheless, *BE* can also be preceded by auxiliaries (*has been, will be*) and semi modals (*is to be, is supposed to be, used to be*, etc.) (Biber, et al., 2002). In the L1-Malay learner sub-corpus, existential *there* clauses are found to be formed mainly with copula *BE* as in (a), auxiliary *BE* as in (b) and modal auxiliary as in (c).

a) There ***are*** different types of censorship; among them are internet, music, television, film, movie and radio censorship. A0004
b) There ***are*** some sort of punishments ***imposed*** by the court like death, whipping, fines, compensation orders and costs and also police supervision. L0032a
c) …there ***will be*** high tendency for the criminals to repeat the crimes… FP0071

Table 4.19 summarises the distribution of *BE* in the formation of existential *there* clauses.

Table 4.19: Distribution of *BE* According to Functions in Existential *there* Clauses in L1-Malay Data

Function	Token	Ratio	%
Copula *BE*	497	2506.78	87.19
Auxiliary *BE*	41	206.80	7.19
Auxiliary modal + *BE*	32	161.40	5.61
Total	**570**	**2874.98**	**100**

As can be seen from Table 4.19, a large majority of the existential *there* constructions are formed with copula *BE* (87.19%). The use of auxiliary *BE* and modal auxiliary in the formation of existential *there* are very restricted, recording about 7.19% and 5.61% respectively. In general, existential *there* clauses in the L1-Malay learner sub-corpus are mainly constructed with copula *BE*.

4.2.2.5.1 Type of Subjects

As explained earlier existential *there* clauses generally take *there* + *BE* + *NP* + *adverbial* structure, hence only the type of subjects was included in the analysis of the constituents surrounding *BE*. Existential *there* structure takes mostly indefinite noun phrase (*some students, nobody, nothing*) as its notional subject. However, definite noun phrase (*three reasons, the factor*) and proper nouns are also used as the notional subjects (Biber et al., 1999). In the L1-Malay learner sub-corpus, nearly all the occurrences of the existential *there* clauses take indefinite NP as the notional subjects as shown in Table 4.20.

Table 4.20: Distribution of Type of Subjects in *BE* in Existential *there* Clauses in L1-Malay Data

	Indefinite Noun Phrase (Ind NP)		Definite Noun Phrase (Def NP)	
Function	**Token**	**%**	**Token**	**%**
Copula *BE*	487	85.44	10	1.75
Auxiliary *BE*	39	6.84	2	0.35
modal + be	32	5.61	0	0.00
Total	**558**	**97.89**	**12**	**2.11**

Approximately 97.89% of the existential *there* clauses are formed with indefinite NP, while only about 2.11% are formed with definite NP. Table 4.20 also shows that 85.44% of the indefinite NP subjects are used in the copula *BE* construction compared to about 6.84% in the auxiliary *BE* construction and 5.61% in *modal + BE* construction. The figures show that the existential *there* clauses in the L1-Malay learner sub-corpus concentrates mainly on one pattern; that is *There + Cop BE + Ind NP + adverbial.*

4.2.2.5.2 Summary of the Patterns of Use of BE in Existential there

The patterns of the use of *BE* in the existential *there* clauses in the L1-Malay learner sub-corpus can be summarised as follows:

1. Both finite *BE* (copula and auxiliary *BE*) and non-finite *BE* (infinitive *be*) are used in the formation of existential *there* construction, but the majority of existential *there* constructions in the sub-corpus are formed with copula *BE.*
2. Nearly all of the existential *there* clauses take indefinite noun phrases as the notional subjects and they are most likely to occur in *There + Cop BE + Ind NP + adverbial* structure.

4.2.2.6 Patterns of the Use of *BE* in *It*-Clefts

Another unique construction that involves the use of *BE* is *it*-cleft. The structure consists of pronoun *it*, a form of *BE*, a focused element which may be a noun phrase, a prepositional phrase, an adverb phrase or adverbial clause and a relative-like dependent clauses introduced by *that, who/which* or zero as in *It is* ***here*** *[that the finite element analysis comes into its own]* (Biber et al., 2002, p. 420). Note that in the sample the focused element is bold and the dependent clause is in square brackets.

Since pronoun *it* is used as the subject, only singular *BE* forms *is, 's* and *was* are used in the formation of *it*-cleft clauses. Table 4.21 summarises the distribution of *BE* in the formation of *it*-cleft clauses.

Table 4.21: Distribution of *BE* According to Forms in *It*-Cleft Clauses in L1-Malay Data

Form	Tokens	%
is	195	88.64
's	19	8.64
was	6	2.73
Total	**220**	**100.00**

The overall percentage of occurrences of *it*-cleft clauses in the L1-Malay learner sub-corpus is only about 3.43% (refer to Table 4.5). *BE* used in the formation of *it*-cleft clauses mainly functions as copula in a main verb position. Approximately 88.64% of the clauses are formed with *is* and only about 8.64% with the contracted form *'s* and 2.73% with the past form *was*.

4.2.2.6.1 Subject Predicates

In order to obtain a better understanding of the *it*-cleft clauses found in the L1-Malay ESL learner data, the analysis of the post-*BE* constituent was carried out. The only constituent analysed is the focused elements (subject predicates). They include noun phrase (NP), adjective phrase (AP), prepositional phrase (PP) and adverbial clause (AdvC). Table 4.22 summarises the findings.

Table 4.22: Distribution of Subject Predicates in *It*-Cleft Clauses in L1-Malay Data

	Noun Phrase (NP)		Adjective Phrase (AP)		Prepositional Phrase (PP)		Adverbial Clause (AdvC)	
Form	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
is	55	25	114	51.82	12	5.45	14	5.45
's	9	4.09	8	3.64	2	0.91	0	0.00
was	4	1.82	2	0.91	0	0.00	0	0.00
Total	**68**	**30.91**	**124**	**56.36**	**14**	**6.36**	**14**	**6.36**

As shown in Table 4.22, AP is used most frequently as complements to *it*-cleft clauses. About 56.36% of *BE* take AP as the subject predicates, while NP predicates record

approximately 30.91% of use. PP and AdvC predicates record a low percentage of approximately 6.36% each. In general, *it*-cleft clauses in the L1-Malay ESL learner data are complemented mainly by AP and NP predicates, with AP predicates occurring more frequently than NP predicates.

4.2.2.6.2 Summary of the Patterns of Use of BE in It-Clefts

The patterns of the use of *BE* in the *it*-cleft clauses can be summarised as:

1. Almost all the *BE* in the *it*-cleft clauses functions as copular.
2. AP predicates are used most frequently in the *it*-cleft clauses followed by NP predicates. *It*-cleft clauses in the L1-Malay ESL learner data are most likely constructed as follows:
    a) *It + BE + AP*
    b) *It + BE + NP*

4.2.3 Patterns of Grammatical Use of Non-Finite *BE*

The investigation of *BE* in this study was widened to include the patterns of the grammatical use of non-finite *BE* forms namely, *be*, *been* and *being*. The findings are presented according to the different forms, functions and in relation to the syntactic environments, which include the type of subjects, subject predicates and class of post-*BE* verbs.

4.2.3.1 Overall Pattern of Grammatical Use of Non-Finite *BE*

Table 4.23 below summarises the tokens and percentages of each non-finite *BE* form. The distribution is also summarised according to the functions performed by each form.

Table 4.23: Distribution of Non-Finite *BE* According to Forms and Functions in L1-Malay Data

Form	Function	Token	%	Overall %
be	future tense (*modal + be*)	489	47.6	37.82
	modal in passive voice (*modal + be + Ven*)	515	50.1	39.83
	modal in progressive aspect (*modal + be + Ving*)	24	2.3	1.86
	Total	**1028**	**100**	**79.51**
been	present/past perfect (*have/has/had+ been*)	12	5.6	0.93
	perfect passive (*have/has/had+ been + Ven*)	181	84.6	14.00
	perfect progressive (*have/has/had + been + Ving*)	21	9.8	1.62
	Total	**214**	**100**	**16.55**
being	progressive passive (*be + being + Ven*)	51	100	3.94
	Total	**51**	**100**	**3.94**

As shown in Table 4.23, infinitive *be* records the highest percentage of overall occurrences of 79.51%. In the L1-Malay learner sub-corpus infinitive *be* is primarily used in the construction events in the future and passive with 47.6% and 50.1% of occurrences respectively.

The overall percentage of *been* in the data is only about 16.55%. As presented in Table 4.23, about 84.6% of *been* are used in the formation of perfect passives, which suggests that *been* in the L1-Malay learner data is used exclusively in the formation of perfect passives. Only 5.6% of *been* are used in the formation of present/past perfect and about 9.8% in the construction of perfect progressive.

The form *being* is mainly used in the formation of progressive passive (*BE + being + Ven*). The overall percentage recorded for *being* in progressive passives is approximately 3.94%, suggesting that its use among the L1-Malay ESL learners is very limited.

4.2.3.2. Type of Subjects

This section presents the findings of the types of subjects preceding non-finite *BE* forms. The subjects are categorised into three categories, namely nouns (NP), personal pronouns (PPN) and pronouns (PN). Pronouns other than PPN, which include definite,

indefinite and *wh*-pronouns are categorised under PN. Table 4.24 summarises the findings.

Table 4.24: Distribution of Type of Subjects in Non-Finite *BE* Constructions in L1-Malay Data

Form	NP		PPN		PN	
	Token	%	Token	%	Token	%
be	600	46.40	356	27.53	82	6.34
been	125	9.67	66	5.10	10	0.77
being	42	3.25	5	0.39	4	0.31
Total	**767**	**59.32**	**427**	**33.02**	**96**	**7.42**

As shown in Table 4.24, NP subjects are used most frequently with all the non-finite *BE* forms with an overall percentage of 59.32%. Approximately 46.40% of the NP subjects occur in infinitive *be* constructions, about 9.67% occur in the construction of perfect aspect (*been*) and about 3.25% in the construction of progressive passive (*being*). PPN records the second highest percentage of overall use of 33.02%, with 27.53% occurring in infinitive *be* constructions, followed by about 5.10% in *been* constructions and only about 0.39% in *being* constructions. Other pronouns (PN) record a low overall percentage of 7.42%. In general, NP subjects are most preferred in the non-finite *BE* constructions followed by PPN subjects.

4.2.3.3 Subject Predicates

This section presents the distribution of subject predicates in the non-finite *BE* constructions. Only infinitive *be* and *been* constructions involve the use of subject predicates in *Subject + modal + be + Complement* (*Money can be **evil***) or *Subject + have + been + Complement* (*It has only been **a dream***). Nevertheless, only the distribution of subject predicates in infinitive *be* constructions is discussed here. There are very few instances of *been* constructions that are complemented by subject predicates and for this reason they are excluded from this analysis. Table 4.25 summarises the distribution of subject predicatives for infinitive *be* constructions.

Table 4.25: Distribution of Subject Predicates in Infinitive *be* Constructions in L1-Malay Data

Form	NP		AP		PP	
	Token	%	Token	%	Token	%
be	196	40.08	210	42.94	73	14.93

Only three types of subject predicates are found to complement infinitive *be* clauses, namely noun phrase (NP), adjective phrase (AP) and prepositional phrase (PP). As shown in Table 4.25, infinitive *be* clauses are mostly complemented by either AP or NP predicates. Both AP and NP predicates record almost similar percentages; 42.94% and 40.08% respectively, while, PP predicates only record about 14.93% of occurrences. In general, infinitive *be* constructions in the L1-Malay learner data are complemented mainly by AP and NP predicates

4.2.3.4 Class of Post-*BE* Verbs

In order to obtain a more comprehensive account of the patterns of use of the non-finite *BE*, the class of post-*BE* verbs was also analysed. The analysis include infinitive *be* constructions in future passives (*modal + be + Ven*) and progressive aspect (*modal + be + Ving*), *been* in perfect passive (*have + been + Ved*) and perfect progressive (*have + been + Ving*) and *being* in progressive passive (*be + being + Ved*). Table 4.26 summarises the findings.

Table 4.26: Distribution of Class of Post-*BE* Verbs in Non-Finite *BE* Constructions in L1-Malay Data

Form	Vt		Uer		Uac	
	Token	%	Token	%	Token	%
be	497	62.75	42	5.30	0	0.00
been	185	23.36	13	1.64	4	0.51
being	46	5.81	5	0.63	0	0.00
Total	**728**	**91.92**	**60**	**7.58**	**4**	**0.51**

As displayed in Table 4.26, Vt verbs record the highest percentage of overall use (91.92%) compared to Uer verbs (7.58%) and Uac verbs (0.51%). The occurrences of Vt verbs are constantly high across all the non-finite *BE* constructions. Approximately

62.75% of Vt verbs are used in the infinitive *be* constructions, another 23.36% in *been* constructions and about 5.81% in *being* constructions. As for Uer and Uac verbs, they are not used as frequently as transitive verbs. In general, it can be summarised that non-finite *BE* constructions mainly occur with transitive verbs.

4.2.3.5 Summary of Patterns of Grammatical Use of Non-Finite *BE*

This section summarises the patterns of the grammatical use of non-finite *BE* forms in the L1-Malay learner sub-corpus.

Infinitive *be*

Infinitive *be* is mainly used to express the future (*modal + be*) and in the formation of passives (*modal + be + Ved*). The *modal + be* structure is most frequently preceded by NP subjects, followed by PPN subjects and are most often complemented by AP and NP predicates, which would likely result in the following constructions:

a) *NP + modal + be + AP*

b) *PPN + modal + be + AP*

c) *NP + modal + be + NP*

d) *PPN + modal + be + NP*

The *modal + be + Ved* structure is also mostly preceded by NP and PPN subjects and followed by Vt verbs. The following constructions summarise the most likely patterns of infinitive *be* in the formation of passives:

e) *NP + modal + be + Vt-ed*

f) *PPN + modal + be + Vt-ed*

Non-finite *been*

In the L1-Malay learner sub-corpus, *been* is mainly used in the formation of perfect passive (*have + been + Ved*). The constructions are often preceded by NP and PPN

subjects and followed by Vt verbs, which most likely would result in the following construction:

a) *NP/PPN + have + been + Vt-ed*

Non-finite *being*

The form *being* is comparatively rare in L1-Malay learner data and it is only used in the formation of progressive passive (*be + being + Ven*). The construction is most often preceded by NP subjects and followed by Vt verbs as shown in (a) below:

a) *NP + be + being + Vt-ed*

4.2.4 Influence of Syntactic Environments on Grammatical Use of *BE*

This section addresses Research Question 4, which focuses on the influence of the syntactic environments on the grammatical use of *BE*.

4.2.4.1 Influence of Syntactic Environments on Grammatical Use of Finite *BE*

In order to determine whether the grammatical use of *BE* is influenced by the syntactic environments, the analysis went further to include the constituents occurring before and after *BE*. The discussion will only focus on *BE* functioning as copular and auxiliary.

4.2.4.1.1 Copula BE

The syntactic environments under investigation for the grammatical use of copula *BE* include the type of subjects and subject predicates.

4.2.4.1.1.1 Type of Subjects

In determining the pattern of the use of *BE* in the L1-Malay learner sub-corpus, all the subjects preceding the grammatical copula *BE* constructions are coded. The analysis includes copula *BE* in four major types of clauses; declaratives, negatives, interrogatives and *it*-clefts. The subjects are divided into noun (NP), personal pronoun (PPN), definite

pronoun (DPN), indefinite pronoun (IPN) and wh-pronoun (QPN). Table 4.27 summarises the findings

Table 4.27: Distribution of Type of Subjects in Copula *BE* Constructions in L1-Malay Data

	Noun (NP)		Personal Pronoun (PPN)		Definite Pronoun (DPN)		Indefinite Pronoun (IPN)		Wh-Pronoun (QPN)	
	Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
Declarative	2727	55.63	975	19.89	368	7.51	50	1.02	141	2.88
Negative	209	4.26	102	2.08	24	0.49	3	0.06	10	0.2
Interrogative	20	0.4	34	0.69	19	0.39	0	0	0	0
It-Cleft	65	1.32	123	2.51	14	0.29	12	2.25	6	0.12
Total	**3021**	**61.61**	**1234**	**25.17**	**425**	**8.68**	**65**	**3.33**	**157**	**3.2**

As can be seen from Table 4.27, copular constructions in the L1-Malay learner data take most frequently NP subjects with 61.61% of occurrences. According to Biber et al. (1999), it is common for nouns to be used more often as subjects in academic essays as learners tend to make constant reference to the essay topics in their strategy to draw focus to the issue being discussed. Academic prose in the Longman Spoken and Written English Corpus contains approximately 280,000pmw of nouns compared to only about 30,000pmw of pronouns (Biber et al., 1999). These figures confirm that it is common for nouns to be used more often than pronouns in academic writings. Nevertheless, it is difficult to associate the grammatical use of copula *BE* with the subjects used as the patterns of use seem to suggest that learners' choice to NP subjects was driven by the need to conform to academic writing convention.

PPN is the second most used subjects (25.17%) in the major *BE* constructions compared to other pronoun categories namely, definite, indefinite and *wh*- pronouns. Learners' preference of PPN lies in the functions that they perform. PPNs generally serve as references to replace noun phrases in texts and they are better suited for that purpose since they can be very specific on one hand and cover a wide range of subjects/items on the other. There are different forms of PPNs that are categorised according to number, person, case and gender making them better choices as references (Biber et al., 1999).

4.2.4.1.1.2 Subject Predicates

Another important constituent that is analysed in the grammatical copula *BE* constructions is the subject predicates complementing *BE*. The analysis includes all the predicates that are used by the learners and they can be divided into two categories; phrasal and clausal predicates. Table 4.28 below summarises the findings.

Table 4.28: Distribution of Subject Predicates in Copula *BE* Constructions in L1-Malay Data

	Noun Phrase (NP)		Adjective Phrase (AP)		Prepositional Phrase (PP)		Infinitive Clause (InfC)		That Clause (ThtC)		Wh-Clause (WhC)		Adverbial Clause (AdvC)	
	Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
Declarative	2041	43.81	1314	28.20	366	7.86	129	2.77	65	1.4	95	2.04	0	0
Negative	140	3.00	171	3.67	27	0.58	6	0.13	4	0.09	0	0	0	0
Interrogative	25	0.54	34	0.73	8	0.17	0	0	0	0	8	0.17	6	0.13
It-Cleft	65	1.40	123	2.64	14	0.30	0	0	6	0.13	0	0	12	0.26
Total	**2271**	**48.74**	**1642**	**35.24**	**415**	**8.91**	**135**	**2.90**	**75**	**1.61**	**103**	**2.21**	**18**	**0.39**

As shown in Table 4.28, only two types of predicates occur most frequently in the L1-Malay ESL learner data and they are NP and AP predicates. The trend is consistent across all major copular constructions. NP occurs more frequently than AP with 48.74% and 35.24% respectively. Lee and Huang (2004) also reported better performance of *BE-noun* and *BE-adjective* in the Chinese learners' data they analysed. According to Lee and Huang (2004) *BE-noun* and *BE-adjective* are comparatively easier for the Chinese learners as the same structures also exist in the Chinese grammar. The level of difficulties of the different copula *BE* constructions, however, does not seem to affect the pattern of use by the Malay learners. In general, L1-Malay learners in this study are competent in all the copular constructions, but some constructions are used more frequently (*BE-noun*, *BE-adjective*) than others (*BE-preposition*) due mainly to the requirement of the academic writing. As pointed out by Biber et al. (1999), the most common predicatives of copular constructions in academic prose are noun predicates and adjective predicates, as they perform specific functions in the academic prose. Noun predicates are used to either characterise or

identify an attribute, while adjective predicates are commonly employed to express specific evaluation. Thus, the high frequency of both NP and AP predicates complementing copula *BE* in the L1-Malay ESL learner data is considered common in academic writing, therefore, not be directly influenced by the types of the subject predicates.

4.2.4.1.2 Auxiliary BE

4.2.4.1.2.1 Type of Subjects

As can be seen in Table 4.29, auxiliary *BE* constructions have also recorded higher use of NP subjects; about 44.19% in the passive constructions and 24.28% in the progressive constructions. The use of NP subjects seems to dominate the grammatical auxiliary *BE* constructions with the overall percentage of about 68%. As mentioned earlier, it is common for NP subjects to be used more frequently in academic prose (Biber et al., 1999), thus, the grammatical use of auxiliary *BE* by the L1-Malay learners in this study could not be directly associated to the types of subjects used.

Table 4.29: Distribution of Type of Subjects in Auxiliary *BE* Constructions in L1-Malay Data

		Noun (NP)		Personal Pronoun (PPN)		Definite Pronoun (DPN)		Indefinite Pronoun (IPN)		Wh-Pronoun (QPN)	
		Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
Aux-Pas	Declarative	680	42.45	203	12.67166042	22	1.37	11	0.69	20	1.25
	Negative	25	1.56	19	1.19	1	0.06	0	0.00	0	0.00
	Interrogative	3	0.19	1	0.06	0	0.00	0	0.00	0	0.00
	Total	**708**	**44.19**	**223**	**13.92**	**23**	**1.44**	**11**	**0.69**	**20**	**1.25**
Aux-Prog	Declarative	375	23.41	163	10.17	4	0.25	8	0.50	39	2.43
	Negative	12	0.75	11	0.69	0	0.00	0	0.00	1	0.06
	Interrogative	2	0.12	1	0.06	1	0.06	0	0.00	0	0.00
	Total	**389**	**24.28**	**175**	**10.92**	**5**	**0.31**	**8**	**0.50**	**40**	**2.50**

4.2.4.1.2.2 Class of Post-*BE* Verbs

Auxiliary *BE* constructions involve the use of post-*BE* lexical verbs. The post-*BE* verbs are classified into three main categories; transitive verbs (Vt), unergative verbs (Uer)

and unaccusative verbs (Uac). Table 4.30 summarises the distribution of post-*BE* verbs in the auxiliary *BE* constructions.

Table 4.30: Distribution of Post-*BE* Verbs in Auxiliary *BE* Constructions in L1-Malay Data

		Transitive (Vt)		Unergative (Uer)		Unaccusative (Uac)	
		Tokens	**%**	**Tokens**	**%**	**Tokens**	**%**
Aux-Pas	Declarative	936	63.93	0	0.0	0	0.0
	Negative	45	3.07	0	0.0	0	0.0
	Interrogative	4	0.27	0	0.0	0	0.0
	Total	**985**	**67.28**	**0**	**0.0**	**0**	**0.0**
Aux-Prog	Declarative	268	18.31	147	10.0	36	2.46
	Negative	16	1.09	8	0.5	0	0.00
	Interrogative	4	0.27	0	0.0	0	0.00
	Total	**288**	**19.67**	**155**	**10.6**	**36**	**2.46**

In the progressive constructions, auxiliary *BE* is found to be proceeded mostly by Vt and Uer verbs. Nevertheless, Vt verbs appear to occur more frequently (19.67%) than Uer verbs (10.6%). As for the passive constructions, as anticipated only Vt verbs are used in this construction (67.28%). In general, for both auxiliary *BE* constructions the use of Vt verbs is more prominent compared to unergative and unaccusative verbs.

The findings show that the learners are aware that passive voice can only be constructed with transitive verbs, which is different with progressive constructions that allow for the use of both transitive and intransitive verbs. Evidently, the learners are aware of the deep structures of both transitive and intransitive verbs. In addition, the learners also exhibit high competency in the intricate process of object-subject inversion that is primary in constructing passives. Nevertheless, it is difficult to determine and measure if the class of post-*BE* verbs influence the grammatical constructions of passive voice and progressive aspect. It can be concluded that the L1-Malay learners in this study are generally competent in the constructions of both progressive aspect and passive voice.

4.2.4.1.3 Summary of the Influence of Syntactic Environments on Grammatical Use of BE

As previously mentioned, the patterns of the grammatical constructions of finite *BE* in the L1-Malay learner data are consistent with that reported in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). According to Biber et al. (1999) it is common for academic prose to contain higher number of NP and PPN subjects and for copula *BE* to be complemented by NP and AP predicates. Since MACLE consists of mainly argumentative essays, the higher frequencies of NP and PPN used as subjects and the frequent occurrences of *BE*-noun and *BE*-adjectives are considered common. The preference learners exhibit to these *BE* structures (e.g. *BE*-noun, *BE*-adjective) is believed to be shaped by the register (written) and writing genre (academic/argumentative), hence, is not determined or influenced by the syntactic environments.

The findings also suggest that frequency of use could not be equated with learners' proficiency. Some structures for instance *BE + Ving* and *BE-preposition* are not very common in the data, as they are less required in academic composition. Progressives for instance are more common in the spoken register than in the written (Biber et al., 1999), while prepositional phrase according to Biber et al. (1999) is the least common type of complement to *BE*. Therefore, the low frequency of these structures is not the outcome of learners' incompetency in them. In general, the L1-Malay learners in this study are competent in all the copula *BE* and auxiliary *BE* structures, which is evident in the grammatical use of these structures in the learner essays.

4.3 Ungrammatical Use of *BE* in the L1-Malay Learner Sub-Corpus

The analysis of the ungrammatical use of *BE* is an integral part of the study as the findings will enable the researcher to draw more comprehensive conclusions of the use of *BE* by the L1-Malay learners in this study. This section focuses on the results of the

ungrammatical use of *BE* in this study. The terms incorrect use, misuse and error are used interchangeably in this chapter to refer to the ungrammatical use.

4.3.1 Distribution of Ungrammatical Use of *BE*

The first part of the analysis involves investigation on the ungrammatical use of all the finite and non-finite *BE*. The ungrammatical use of *BE* are compared to the grammatical use to gauge the learners' overall performance.

4.3.1.1 Distribution of Ungrammatical Use of *BE* According to Forms

Table 4.31 summarises the distribution of grammatical and ungrammatical uses of *BE* according to forms.

Table 4.31: Grammatical and Ungrammatical Uses of *BE* According to Forms in L1-Malay Data

	Grammatical		Ungrammatical		
Form	**Token**	**%**	**Token**	**%**	**Rate (%)**
am	40	0.44	0	0.00	0.00
is	4039	44.16	307	3.36	7
's	137	1.50	0	0.00	0.00
isn't	0	0.00	0	0.00	0.00
are	2173	23.76	399	4.36	16
're	12	0.13	0	0.00	0.00
aren't	0	0.00	0	0.00	0.00
was	345	3.77	78	0.85	18
wasn't	0	0.00	0	0.00	0.00
were	215	2.35	31	0.34	14
weren't	0	0.00	0	0.00	0.00
be	1028	11.24	66	0.72	6
been	214	2.34	10	0.11	4
being	51	0.56	0	0.00	0.00
Total	**8254**	**90.25**	**891**	**9.74**	**9.74**

Table 4.31 shows that only 9.74% of *BE* are incorrectly used and they include *are, is, was, were, be* and *been.* Other forms record no occurrences of ungrammatical use. The present forms (*is, are*) record higher combined percentage of ungrammatical use of 7.7% compared to the past forms (*was, were*), which record the combined percentage of only 1.2%. In comparison to the finite forms, which have recorded the overall

percentage of ungrammatical use of 8.91%, non-finite forms recorded relatively fewer instances of ungrammatical use of only 0.83%.

In order to gauge the real extent of learners' difficulties with each *BE* form, the rate of the ungrammatical use was calculated. The calculation involves the tokens of ungrammatical use of each form divided with the total tokens of use (combination of grammatical and ungrammatical uses) of the same form and multiplied by 100 to obtain the percentiles. From this calculation, it is found that learners tend to make more errors with *was, are* and *were*, which record the rates of 18%, 16% and 14% respectively. Even though, the past forms *was* and *were* are not as frequently used as the present forms *is* and *are*, they tend to be incorrectly used more often by the learners. This finding suggests that some learners are still having difficulty with tense feature. The tendency to misuse the plural forms *are* and *were* also suggests that marking agreement is also a problem to some learners.

The figures presented in Table 4.31 suggest that in general most learners do not face much difficulty in using *BE* as evident in the low percentages of the ungrammatical use across all forms. In addition, the findings also reveal an interesting aspect of L1-Malay learners' competency in use of *BE* that is some learners are still struggling with the fundamental projection of agreement and tense.

4.3.1.2 Distribution of Ungrammatical Use of BE According to Functions

Table 4.32 summarises the ungrammatical use of finite *BE* forms according to the functions they perform. The tokens and percentages of the grammatical use are also presented for comparison purposes.

Table 4.32: Grammatical and Ungrammatical Uses of Finite *BE* According to Functions in L1-Malay Data

	Grammatical		Ungrammatical		
Function	**Token**	**%**	**Token**	**%**	**Rate (%)**
Copula	4287	57.1	266	3.5	6.2
Auxiliary	1387	18.5	218	2.9	15.7
Negative operator	424	5.6	16	0.2	3.8
Interrogative operator	81	1.1	4	0.1	4.9
Existential *there*	544	7.2	31	0.4	5.7
It-cleft	239	3.2	16	0.2	6.7
Total	**6962**	**92.7**	**551**	**7.3**	**7.3**

As shown in Table 4.32, the ungrammatical use of *BE* according to functions is significantly low (7.3%) compared to the grammatical use (92.7%). Copular constructions record the highest percentage of ungrammatical use of 3.5%, followed by auxiliary constructions, which record a percentage of 2.9%. Nevertheless, the rates of ungrammatical use provide a different scenario; auxiliary *BE* is found to be used incorrectly more often (15.7%) than copula *BE* (6.2%). The finding suggests that for some learners the use *BE* as an auxiliary is more challenging than as a main verb.

4.3.2 Patterns of the Ungrammatical Use of *BE*

The findings for the patterns of the ungrammatical use of *BE* are presented according the major types of ungrammatical use (Section 4.3.2.1), followed by the findings for the patterns of the ungrammatical use (Section 4.3.2.2).

4.3.2.1 Major Types of Ungrammatical Use of *BE*

The analysis of the ungrammatical use of *BE* is based on the analysis parameters set earlier in the study. The parameters is set based on four major types of misuse of *BE* already attested in previous studies, which are agreement (Agr), tense (Tns), overgeneration (Ovg) and omission (Null). The samples for each type of ungrammatical use are shown below:

Agr: **If the motive <u>are</u> pure then the fruit can be very good.* K0084

Tns: **This kind of thing in the past <u>is</u> just a dream.* F0106

Ovg: **This <u>is always happen</u> to the people which is new with the computer Technologies.* C0007

Null: **Co-cirricular activities <u>[are]</u> also important to the student for get a job.* B0039-05

Table 4.33 below summarises the tokens and percentages of the major ungrammatical use according to *BE* forms.

Table 4.33: Distribution of Ungrammatical Use of *BE* According to Forms in L1-Malay Data

	Agreement		Tense		Overgeneration		Omission		Others	
	Agr		Tns		Ovg		Null			
	Tokens	%	Tokens	%	Tokens	%	Tokens	%	Tokens	%
is	59	6.62	21	2.36	144	16.16	81	9.09	2	0.22
are	111	12.46	22	2.47	151	16.95	110	12.35	5	0.56
was	12	1.35	32	3.59	30	3.37	0	0.00	4	0.45
were	16	1.80	10	1.12	5	0.56	0	0.00	0	0.00
be	0	0.00	0	0.00	55	6.17	11	1.23	0	0.00
been	0	0.00	0	0.00	10	1.12	0	0.00	0	0.00
Total	**198**	**22.22**	**85**	**9.54**	**395**	**44.33**	**202**	**22.67**	**11**	**1.23**

As shown in Table 4.33, overgeneration records the highest percentage of ungrammatical use. Approximately 44.33% of the total ungrammatical use of *BE* comprise overgeneration errors. There are also fairly high occurrences of omission and agreement errors, 22.67% and 22.22% respectively. Errors in tense, however, record a very low percentage of about 9.54%. These figures show that overgeneration, omission and agreement are the major types of errors found in the L1-Malay learner sub-corpus. In terms of ranks, the ungrammatical use can be ranked as Ovg>Null>Agr>Tns.

In terms of finiteness, it is found that finite *BE* forms are misused more frequently than non-finite forms. Non-finite forms are only found to occur in overgeneration and omission instances. Since non-finite *BE* forms do not carry inflectional features (agreement and tense), there are no agreement and tense errors of non-finite *BE* recorded in the data.

The ungrammatical use of finite *BE* are also analysed in terms of the two major functions they perform; copular and auxiliary. Overgeneration is not included in this analysis since the function of the overgenerated *BE* cannot be clearly determined. The summary of the results is displayed in Table 4.34.

Table 4.34: Distribution of Ungrammatical Use of Finite *BE* According to Functions in L1-Malay Data

	Copular		Auxiliary	
	Tokens	**%**	**Tokens**	**%**
Agreement	141	27.27	57	11.78
Tense	44	9.09	41	8.47
Omission	81	16.74	109	22.52
Others	0	0.00	11	2.27
Total	**266**	**54.96**	**218**	**45.04**

It is found that copula *BE* tends to be ungrammatically used slightly more frequently than auxiliary *BE*; 54.96% and 45.04% respectively. Copula *BE* records higher percentage of agreement errors than auxiliary *BE*; 27.27% and 11.78% respectively. The opposite result is obtained for omission, auxiliary *BE* records 22.52% of omissions, while copula *BE* records 16.74% of omissions. Nevertheless, there is not much difference in the percentages recorded for tense errors for both copula *BE* and auxiliary *BE* constructions; 9.09% and 8.47% respectively.

In sum, the results of the ungrammatical use of *BE* indicate firstly, that overgeneration is a major problem among the L1-Malay ESL learners, followed by omission and agreement. Secondly, learners have the tendency to misuse finite *BE* more frequently than non-finite *BE*. Thirdly, the learners also face problems with the correct use of both copula and auxiliary *BE*, but the rates of occurrences in Table 4.32 suggest that auxiliary *BE* constructions tend to be more problematic to the learners than copula *BE* constructions.

4.3.2.2 Patterns of *BE* in the Major Types of Ungrammatical Use

In order to extract the patterns of *BE* in each type of the ungrammatical use, the errors are analysed in relation to the syntactic environments. In doing so, each type of ungrammatical use is analysed separately. The following sections present the results from the in-depth analyses of the two major types of ungrammatical use of *BE* in this study; overgeneration and omission.

4.3.2.2.1 Overgeneration of BE

Previous studies (Arshad & Hawanum, 2010; Balcom, 1997; Fleta, 2003; Hirakawa, 2006; Ionin & Wexler, 2001; Ju, 2000; Lee & Huang, 2004; Oshita, 2000; Park & Lakshmanan, 2007; Ting et al., 2010; Wee, Sim & Kamaruzam, 2010; Wee, 2009; Yip, 1994) highlighted a unique construction where *BE* is inserted in a non-obligatory context and combined with a main verb to produce *be + V* as in '*the lion **is go** down*' (Ionin & Wexler, 2001, p. 110). This section reports these constructions which are termed as overgeneration of *BE*; a term used by Ionin and Wexler (2001) to describe insertion of *BE* before a main verb as a mechanism to check tense/agreement feature. The analyses are conducted according to the forms and types of the lexical verbs. The forms of post-*BE* verbs include the base form -*V* (*eat, talk*), third person singular present tense form -*Vs* (*eats, talks*), past form –*Ved* (*talked*) and past participle form *Ven* (*eaten*). Since participle -*ed* is used for both past tense and past participle of regular verbs and there are not many cases of irregular verbs found in overgeneration constructions, *Ved* is used to represent the past tense as well as the past participle form in this study. The class of lexical verbs are categorised into transitive verb (Vt), unergative (Uer) and unaccusative (Uac) following the classification proposed by Perlmutter (1978) and Burzio (1986). The results are presented according to finiteness and the syntactic environments.

4.3.2.2.1.1 Overgeneration of *BE* According to Finiteness

This section presents the results of overgeneration occurrences of finite and non-finite *BE* forms. Table 4.35 below summarises the findings.

Table 4.35: Distribution of Overgeneration of *BE* According to Finiteness in L1-Malay Data

	Finite *BE*			**Non-finite *BE***	
	Token	**%**		**Token**	**%**
is	144	36.46	be	55	13.92
are	151	38.23	been	10	2.53
was	30	7.59			
were	5	1.27			
Total	**330**	**83.54**		**65**	**16.46**

As shown in Table 4.35, there is a higher percentage of finite *BE* being overgenerated compared to non-finite *BE*. Approximately 83.54% of the overgeneration instances involve finite forms compared to only about 16.46% involving non-finite forms.

4.3.2.2.1.1.1 Overgeneration of Finite BE

This section presents the findings of overgeneration of finite *BE* in relation to the syntactic environments, which include the form and class of post-*BE* verbs, type of subjects and the presence of intensifiers and auxiliaries.

4.3.2.2.1.1.1.1 Form and Class of Post-BE Verbs

Table 4.36 summarises the results of overgeneration cases according to the form and class of post-*BE* verbs.

Table 4.36: Distribution of Overgeneration of Finite *BE* According to Form and Class of Post-*BE* Verbs in L1-Malay Data

	Tokens	**%**
Vt	129	68.25
Vt-s	13	6.88
Vt-ed	45	23.81
Total Vt	**189**	**100.00**
Uer	52	57.14
Uer-s	14	15.38
Uer-ed	16	17.58
Total Uer	**91**	**100.00**
Uac	25	50.00
Uac-s	4	8.00
Uac-ed	21	42.00
Total Uac	**50**	**100.00**

As shown in Table 4.36, *BE* tends to be overgenerated before the base form verbs and the pattern is consistent across the verb classes; 68.25% (Vt), 57.14% (Uer) and 50% (Uac). There are also some instances of overgeneration occurring before verbs inflected with past tense/past participle *–ed*; 23.81% (Vt), 17.58% (Uer) and 42% (Uac). Between the three verb classes, unaccusatives appear to record the highest overgeneration cases involving participle *–ed.* As for 3rd person singular *–s*, it is in general the least form occurring in the overgeneration cases.

The finding suggests that overgeneration involving transitive and unergative verbs occur more frequently with uninflected verbs. To illustrate this finding more clearly, the overgeneration instances are analysed according to the uninflected and inflected categories. Table 4.37 summarises the results.

Table 4.37: Distribution of Overgeneration of Finite *BE* According to the Forms of Post-*BE* Verbs in L1-Malay Data

	Uninflected Verb		**Inflected Verb**	
	Tokens	**%**	**Tokens**	**%**
Vt	129	39.09	60	18.18
Uer	52	15.76	39	11.82
Uac	25	7.58	25	7.58
Total	**206**	**62.42**	**124**	**37.58**

As shown in Table 4.37, overgeneration of transitive and unergative verbs occurs more frequently with uninflected verbs; 39.09% and 15.76% respectively. The same percentage (7.58%) is recorded for overgeneration with uninflected unaccusative verbs and those inflected. In general, there is a higher overall percentage of overgeneration involving uninflected verbs (62.42%) than those involving inflected verbs (37.58%) across the verb classes, resulting in *BE + V > BE + Ved.* As displayed in Table 4.37, transitive verbs occur most frequently in overgeneration constructions, followed by unergative and unaccusative verbs.

In general, overgeneration of finite *BE* in the L1-Malay learner sub-corpus occurs more frequently with base form verbs and followed by past tense/past participle form and

finally with third person singular form (*V > Ved >Vs*). As for class of verb, transitive verbs are used most frequently, followed by unergative verbs and unaccusative verbs (*Vt>Uer>Uac*).

4.3.2.2.1.1.1.2 Type of Subjects

This section presents the results of overgeneration of finite *BE* in relation to the types of subjects preceding *BE*. The subjects are categorised into lexical noun subjects (NP) and pronoun subjects (PN). Table 4.38 summarises the results.

Table 4.38: Distribution of Overgeneration of Finite *BE* According to Type of Subjects in L1-Malay Data

	Noun Subject NP		**Pronoun Subject PN**	
	Token	%	Token	%
is	69	20.91	75	22.73
are	63	19.09	88	26.67
was	15	4.55	15	4.55
were	4	1.21	1	0.30
Total	**151**	**45.76**	**179**	**54.24**

As displayed in Table 4.38, overgeneration occurs more often after PN subjects compared to NP subjects. Nevertheless, the difference in the occurrence of the two types of subjects is fairly small (8.48%), suggesting that there is a general tendency for overgeneration of finite *BE* to occur after both types of subjects.

4.3.2.2.1.1.1.3 Presence of Modal Auxiliaries and Intensifiers

This section presents the results of overgeneration occurrences in the presence of modal auxiliaries and intensifiers. The modal auxiliaries refer to modal verbs such as *will, would, should, can, could, have to* or *ought to,* while the intensifiers refer specifically to degree adverbs such as *always, also* or *only*. Table 4.39 summarises the findings.

Table 4.39: Distribution of Overgeneration of Finite *BE* According to the Presence of Modal Auxiliaries and Intensifiers in L1-Malay Data

	Tokens	**%**
Intensifiers (Degree Adverbs)	70	21.21
Modal Auxiliaries	23	6.97
Total	**93**	**28.18**

As presented in Table 4.39, the overgeneration of finite *BE* forms in the presence of intensifiers and modal auxiliaries is fairly low; 21.21% and 6.97% respectively. The degree adverbs that often appear in the overgeneration constructions include *always, also* and *only*, while only the modals *can* and *cannot* occur in the overgeneration instances. Nevertheless, it is important to note that only about 28% of the overgeneration cases are in the presence of either modal auxiliaries or intensifiers, the other 72% occur without these constituents suggesting that overgeneration of finite *BE* is not affected by the presence of either modal auxiliaries or intensifiers.

4.3.2.2.1.1.1.4 Summary of Patterns of Overgeneration of Finite BE

The patterns of the overgeneration of finite *BE* in the L1-Malay learner sub-corpus can be summarised as:

1. Overgeneration instances of finite *BE* forms are constructed more often with uninflected verbs than inflected verbs *BE + V > BE + Ved*.
2. Overgeneration occurs more frequently with transitive verbs followed by unergative verbs and the least with unaccusative verbs (Vt>Uer>Uac).
3. Pronoun subjects (PN) occur more often before overgeneration of finite *BE* compared to noun subjects (NP).
4. The patterns of finite *BE* overgeneration are more likely to be constructed as the followings:
    i. *PN/NP + BE +V*
    ii. *PN/NP + BE +V > PN/NP + BE +Ved*

iii. *PN/NP +BE +Vt > PN/NP + BE +Uer > PN/NP + BE +Uac*

4.3.2.2.1.2 Overgeneration of Non-finite *BE*

Overgeneration instances involving non-finite *BE* forms has not been reported and discussed in previous research thus far. Despite having comparatively low percentage of occurrences, non-finite *BE* overgeneration was analysed in-depth in this section. The analyses included identifying and quantifying the pre-*BE* and post-*BE* constituents, which include form and class of pre-*BE* and post-*BE* verbs and type of subjects. Intensifiers were generally not present in non-finite *BE* overgeneration, thus, the presence of intensifiers was not included in the analysis.

4.3.2.2.1.2.1 Type of Pre-BE Verbs

Overgeneration of non-finite *BE* occurs in a string of verbs, which comprises of an auxiliary followed by a non-finite *BE* and/or a lexical verb as in *Aux + BE + V*. The auxiliaries that occur in the overgeneration include auxiliary *BE* (*BE + be*) as in (a), modal verb (*modal + be + V*) as in (b) and auxiliary have (*have + be + V*) as in (c).

a) ...and it ***is be*** useful if the student, try to use this chance to make them better...E0045
b) So, they ***should be collect*** many of money until the money make one of root of all evil. F0094
c) ... every countries in the world ***have been started*** to join the campaign against terrorists. K0011

When overgeneration occurs with an auxiliary *BE* as in (a) the verb phrase only contains a string of two verbs and tends to be followed by a subject predicate, which is similar to a copular construction. Whereas, when non-finite *BE* is preceded by either a modal verb or a helping verb *have,* the non-finite *BE* would be followed by a lexical verb as in (b) and (c). The summary of the analysis of the types of pre-*BE* verbs is displayed in Table 4.40.

Table 4.40: Distribution of Overgeneration of Non-Finite *BE* According to Type of Pre-*BE* Verbs in L1-Malay Data

	BE		**have**		**modal**	
	Token	**%**	**Token**	**%**	**Token**	**%**
be	6	9.23	0	0.00	49	75.38
been	1	1.54	9	13.85	0	0.00
Total	**7**	**10.77**	**9**	**13.85**	**49**	**75.38**

As shown in Table 4.40, infinitive *be* are only overgenerated with auxiliary *BE* and modal auxiliaries. Modal auxiliaries are found to occur more frequently than auxiliary *BE*, 75.38% and 9.23% respectively, while overgenerated *been* occurs most frequently with auxiliary verb *have* with 13.85% of occurrences.

Overgeneration of both infinitive *be* and *been* seem to be result of misapplication of the grammar rules. In infinitive *be* overgeneration construction, the modals appear to be used to express permission, ability, obligatory or necessity that are normally formed in *modal + V* construction (*should collect, would buy, can go*). The insertion of infinitive *be* after the modal verb seems to be the result of misapplication of the rules of simple future tense (*modal + be*) as in "*I **will be** there soon"* or passive voice (*modal + be + PP*) as in "*He **could be done** that earlier"*.

Overgeneration of *been*, as mentioned earlier only occurs with auxiliary *have* as in *have + been + V* construction, which has similar construction to perfect passive *have + been + Ved* as in "*have been called"* and "*had been thrown"*. This overgeneration construction seems to be the outcome of misapplication of perfect passive rules to what intended to be present/past perfect. As shown in sample (c) above, "*have been started"* could probably be intended for "*have started"*. The learners may have overgeneralised the perfect passive rule or have mistakenly concluded that every auxiliary *have* needs to be preceded by *been.* In other words, the learners are unable to differentiate the present perfect from the perfect passive and perhaps are unaware of the different functions each performs.

4.3.2.2.1.2.2 Form and Class of Post-BE Verbs

This section presents the results of the form and class of post-*BE* verbs in the non-finite *BE* overgeneration. Table 4.41 summarises the findings.

Table 4.41: Distribution of Overgeneration of Non-Finite *BE* According to Form and Class of Post-*BE* Verbs in L1-Malay Data

	be		**been**	
	Token	**%**	**Token**	**%**
Vt	18	27.69	1	1.54
Vt-s	0	0.00	0	0.00
Vt-ed	12	18.46	6	9.23
Total Vt	**30**	**46.15**	**7**	**10.77**
Uer	8	12.31	0	0.00
Uer-s	0	0.00	0	0.00
Uer-ed	1	1.54	1	1.54
Total Uer	**9**	**13.85**	**1**	**1.54**
Uac	9	13.85	2	3.08
Uac-s	0	0.00	0	0.00
Uac-ed	3	4.62	1	1.54
Total Uac	**12**	**18.46**	**3**	**4.62**

As displayed in Table 4.41, overgeneration instances involving infinitive *be* occur most frequently with the uninflected lexical verbs (*Vt*-27.69%, *Uer*-12.31%, *Uac*-13.85%) and some with verbs inflected with participle *–ed* (Vt-*ed*-18.46%, Uer-*ed*-1.54%, Uac-*ed*-4.62%) and none with verbs inflected with 3rd person singular *–s*. As for overgeneration of *been*, there are very few instances of overgeneration involving the form to produce any significant result.

In terms of the class of post-*BE* verbs, overgeneration instances are found to occur most frequently with transitive verbs (56.92%), followed by with unaccusative verbs (23.08%) and the least with unergative verbs (15.39%). In general, transitive verbs occur most frequently with overgeneration of both infinitive *be* and *been*.

4.3.2.2.1.2.3 Type of Subjects

This section presents the results for the analysis of the types of subjects in the non-finite *BE* overgeneration. Table 4.42 summarises the findings.

Table 4.42: Distribution of Overgeneration of Non-Finite *BE* According to Type of Subjects in L1-Malay Data

	Noun Subject NP		Pronoun Subject PN	
	Token	**%**	**Token**	**%**
be	31	47.69	24	36.92
been	7	10.77	3	4.62
Total	**38**	**58.46**	**27**	**41.54**

Noun subjects appear to occur more frequently (58.46%) in both infinitive *be* and *been* overgeneration constructions, while pronoun subjects record slightly lower percentage of occurrences of 41.54%.

4.3.2.2.1.2.4 Summary of the Patterns of Non-Finite BE Overgeneration

The overgeneration patterns for non-finite *BE* can be summarised as:

1. Overgeneration of infinitive *be* occurs more frequently with modal auxiliaries, uninflected transitive verbs and after NP subjects. The patterns of infinitive *be* overgeneration are as shown below:

    i. *NP + modal + be + Vt > PN + modal + be + Vt*

    ii. *NP/PN + modal + be + Vt> NP/PN + modal + be + Ved*

    iii. *NP/PN + modal + be + Vt> NP/PN + modal + be + Uer> NP/PN + modal + be + Uac*

2. Overgeneration of *been* occurs most frequently with auxiliary verb *have* and more often followed by inflected lexical verbs. The lexical verbs are often inflected with participle *–ed.* NP subjects occur more frequently in overgeneration involving *been*. In terms of the class of post-*BE* verbs, transitive verbs occur most frequently followed by unaccusative and unergative verbs. The patterns of the overgeneration of *been* are as shown below:

    i. *NP + have + been + Ved > PN + have + been + Ved*

    ii. *NP/PN + have + been + Ved > NP/PN + have + been + V*

iii. *NP/PN + have + been + Ved > NP/PN + have + been + Uac-ed > NP/PN + have + been + Uer-ed*

4.3.2.2.2 Omissions of BE

This section presents the findings of the patterns of *BE* omissions in the L1-Malay learner sub-corpus. It is divided into two major sub-sections. The first sub-section presents the findings of copular omission as in (a) and the second sub-section presents the findings of auxiliary omission as in (b) and (c).

a) *They **Ø** afraid to talk in front the public and make others not confident with them. B0038-05
b) *The poor family **Ø always facing** that problems and this make them living in evil. E0014-05
c) *... disagree that our dreams and imagination **Ø taken** away by modernization. K0013

4.3.2.2.2.1 Copula *BE* Omissions

This section presents the results of copular omissions with respect to the syntactic environments, which include the type of subjects, subject predicates, presence of modal auxiliaries and intensifiers.

4.3.2.2.2.1.1 Type of Subjects

Previous research has reported that copula *BE* omission may be influenced by the subject preceding it (Ellis, 1988; Herat, 2005). Ellis (1988) claimed that ESL learners tend to omit copula *BE* more often after a lexical noun subjects (NP) compared to after pronominal subjects (PN). Herat (2005) in his examination of zero *BE* in the Sri Lankan English has also made the same observation. In order to determine the extent of the influence of the type of subjects on *BE* omission cases in the L1-Malay ESL learner data, this study has also examined the type of subjects preceding null *BE* in the L1-Malay learner data. Table 4.43 summarises the findings.

Table 4.43: Distribution of Omission of Copula *BE* According to Type of Subjects in L1-Malay Data

	Noun Subject NP		Pronoun Subject PN	
	Token	**%**	**Token**	**%**
is	30	37.04	21	25.93
are	18	22.22	12	14.81
Total	**48**	**59.26**	**33**	**40.74**

The figures in Table 4.43 show that copula *BE* is omitted more often after NP subjects than after PN subjects. Approximately 59.26% of the omission cases occur after NP subjects compared to about 40.74% after PN subjects. These figures confirm the findings from previous studies that omissions tend to occur more frequently after NP subjects compared to PN subjects (Herat, 2005; Ellis, 1998). The finding suggests that the type of subjects may have some influence over copula *BE* omissions.

4.3.2.2.2.1.2 Subject Predicates

The types of subject predicates complementing the omitted copula *BE* in the L1-Malay learner sub-corpus include noun phrase (NP), adjective phrase (AP), prepositional phrase (PP) and *Wh*-Clause (WhC). Table 4.44 summarises the findings.

Table 4.44: Distribution of Omission of Copula *BE* According to Subject Predicates in L1-Malay Data

	Noun Phrase NP		Adjective Phrase AP		Prepositional Phrase PP		*Wh*-Clause WhC	
	Token	**%**	**Token**	**%**	**Token**	**%**	**Token**	**%**
is	14	17.28	26	32.10	6	7.41	5	6.17
are	7	8.64	23	28.40	0	0.00	0	0.00
Total	**21**	**25.93**	**49**	**60.49**	**6**	**7.41**	**5**	**6.17**

As can be seen in Table 4.44 the omissions of copula *BE* occur mostly before AP predicates (60.49%) and NP predicates (25.93%). Omission cases preceding prepositional phrase and *wh*-clause are very limited; 7.41% and 6.17% respectively. In general, the figures suggest that L1-Malay ESL learners tend to be constrained mostly by *BE-adjective* and *BE-noun* constructions. Platt and Weber (1980), who analysed

spoken Malaysian English also reported the same omission pattern; whereby *BE* was often absent before nominal and adjectival predicates.

Although the pattern of omissions following the types of predicates by the learners in this study is similar to that found by Platt and Weber (1980), the degree of influence each predicate has over the omission instances differs. This degree, however, was not discussed in Platt and Weber (1980). As can be seen from the figures in Table 4.44, even though L1-Malay learners tend to omit *BE* before AP and NP predicates, AP predicates appear to carry stronger influence over null *BE* than NP predicates.

This finding is also consistent to that of Herat (2005) and Unlu and Hatipoglu (2012) who also reported instances of *BE* absence before AP predicates. The findings from previous studies seem to suggest strong *BE-adjective* constraint and that *BE-noun* is easier for learners to use correctly. These findings are consistent with the finding of the current study which also records lesser occurrences of omission before NP predicates.

4.3.2.2.2.1.3 Presence of Intensifiers and Modal Auxiliaries

Previous studies have also reported the tendency L2 learners have to omit copula *BE* in the presence of intensifiers and modal auxiliaries. L1-Chinese learners in the study conducted by Lee and Huang (2004) were reported to omit copula *BE* in the structure of *BE + very/so/not + adjective*, where negator *not* and degree adverbs such as *very* or *so* are used to modify the adjectival complement. Another study investigating L1-Chinese learner language data by Chan (2004) reported on copular omissions after modal auxiliaries. These aspects of copular omissions are further investigated in this study and Table 4.45 summarises the findings.

Table 4.45: Distribution of Omission of Copula *BE* in the Presence of Modal Auxiliaries and Intensifiers in L1-Malay Data

	Tokens	**%**
Negator (Not)	18	22.22
Degree Adverbs	36	44.44
Modal Auxiliaries	5	6.17
Total	**59**	**72.84**

As shown in Table 4.45, most of the omission cases occur in the presence of intensifiers (67%): about 22.22% of which are omissions before the negator *not* and approximately 44.44% before degree adverbs such as *always, very, so*, or *only*. As for omissions after modal auxiliaries, they appear to be very minimal (6.17%) in L1-Malay ESL learner data. The presence of degree adverbs appears to exert a strong influence over copular omissions compared to the presence of negator *not.* This finding suggests that in general intensifiers do pose a strong constraint to copular production among L1-Malay ESL learners. Nevertheless, contrary to the Chinese learners in Chan (2004), L1-Malay learners do not appear to find the presence of modal auxiliaries as problematic.

4.3.2.2.2.1.4 Summary of the Patterns of Copula BE Omission

The patterns of copula *BE* omissions in the L1-Malay learner sub-corpus can be summarised as:

1. Copula *BE* is omitted most frequently before AP predicates followed by NP predicates and seldom omitted before PP and WhC predicates.
2. Omission occurs more frequently after NP subjects compared to after PN subjects.
3. Omission occurs more frequently in the presence of intensifiers but not in the presence of modal auxiliaries.
4. The patterns of copula *BE* omissions are as shown below:

    i. *NP + Ø + AP/NP > PN + Ø + AP/NP*

    ii. *NP/PN + Ø + AP > NP/PN + Ø + NP*

4.3.2.2.2.2 Auxiliary *BE* Omissions

This section presents the results of auxiliary *BE* omissions with respect to the syntactic environments, which include the form and class of post-*BE* verbs, the type of subjects and the presence of intensifiers and modal auxiliaries.

4.3.2.2.2.2.1 Class of Post-BE Verbs

Post-*BE* verbs are found by previous research to have some influence over L2 learners variable use of *BE*. Researchers like Yip (1994), Oshita (2000) and Ju (2000) found that instances of unaccusative verbs being overgenerated to form ungrammatical *BE + Ven* structure as in *was happened* or *was sank*. This finding led the researchers to claim that there is interaction between the class of post-*BE* verbs and the variability in the use of *BE*. Although, there has not been any report on the influence of post-*BE* verbs in the omissions of auxiliary *BE*, this study takes a step further to investigate this aspect. The analysis of post-*BE* verbs in the auxiliary *BE* omission cases was conducted to ascertain if the class of post-*BE* verbs has any influence on omission instances. The post-*BE* verbs are categorised into transitive (Vt), unergative (Uer) and unaccusative (Uac) following the classification proposed by Perlmutter (1978) and Burzio (1986). The findings are also categorised into the functions performed by auxiliary *BE*: auxiliary progressive or auxiliary passive. Table 4.46 below summarises the findings.

Table 4.46: Distribution of Omission of Auxiliary *BE* According to Class of Post-*BE* Verbs in L1-Malay Data

	Transitive Vt		**Unergative Uer**		**Unaccusative Uac**	
	Tokens	**%**	**Tokens**	**%**	**Tokens**	**%**
Aux-Progressive	48	44.04	26	23.85	1	0.92
Aux-Passive	34	31.90	0	0.00	0	0.00
Total	**80**	**75.94**	**28**	**23.85**	**1**	**0.92**

A large majority of auxiliary *BE* omissions occur before transitive verbs (75.94%) and some occur before unergative verbs (23.85%). As for auxiliary *BE* omissions before unaccusative verbs, they are nearly non-existent (0.92%). Omission cases are also

found to involve mostly *BE* in progressive aspect (68.81%) compared to that in the passive voice (31.90%). In the progressive aspect, *BE* is omitted twice more often before transitive verbs (44.04%) than before unergative verbs (23.85%). As for the passive constructions, *BE* is found to be absent only before transitive verbs. In general, *BE* tends to be omitted more frequently in the progressive aspect and involves mostly transitive verbs.

4.3.2.2.2.2.2 Type of Subjects

Similar to copular omissions, auxiliary *BE* omissions are also examined according to the types of subjects preceding them. The summary of the findings are displayed in Table 4.47.

Table 4.47: Distribution of Omission of Auxiliary *BE* According to Type of Subjects in L1-Malay Data

	Noun Subject NP		**Pronoun Subject PN**	
	Tokens	**%**	**Tokens**	**%**
is	21	19.27	8	7.34
are	50	45.87	30	27.52
Total	**71**	**65.14**	**38**	**34.86**

As shown in Table 4.47, omissions of auxiliary *BE* occur more frequently after NP subjects. Approximately 65.14% of the omission occurrences occur after NP subjects, while about 34.86% occur after PN subjects. There seems to be a consistent trend in the omissions of *BE* with regard to the types of subjects. Omissions of both copula *BE* and auxiliary *BE* tend to occur more frequently after NP subjects than after PN subjects.

4.3.2.2.2.2.3 Presence of Intensifiers and Modal Auxiliaries

Omission of auxiliary *BE* is also examined in relation to the presence of intensifiers and also modal auxiliaries. Table 4.48 summarises the findings.

Table 4.48: Distribution of Omission of Auxiliary *BE* in the Presence of Intensifiers and Modal Auxiliaries in L1-Malay Data

	Tokens	%
Negator (Not)	6	5.50
Degree Adverbs	23	21.10
Modal Auxiliaries	6	5.50
Total	**35**	**32.11**

As shown in Table 4.48, the percentage of auxiliary *BE* omissions in the presence of intensifiers and modal auxiliaries are considerably low (32.11%) compared to the omission cases without the presence of these two constituents (67.9%). Out of the 32.11%, about 21.10% occur before degree adverbs such as *always, still* or *so*, while only about 5.5% occur in the presence of negator *not* and the same percentage is also recorded for omissions in the presence of modal auxiliaries. The finding suggests that omissions of auxiliary *BE* tend to be sensitive to the presence of degree adverbs. Nevertheless, the low percentage of omissions in the presence of these constituents (32.11%), suggests that the influence may be very minimal.

4.3.2.2.2.2.4 Summary of the Patterns of Auxiliary BE Omission

The patterns of auxiliary *BE* omissions in the L1-Malay learner sub-corpus can be summarised as:

1. Auxiliary *BE* tends to be omitted most frequently before transitive verbs, followed by unergative verbs.
2. Auxiliary *BE* omissions occur more frequently after NP subjects.
3. The presence of intensifiers and modal auxiliaries has very limited influence on auxiliary *BE* omissions.
4. Auxiliary *BE* omissions tend occur more frequently in the progressive aspect than in the passive voice.
5. The patterns of auxiliary *BE* omissions in the L1-Malay learner sub-corpus are as shown below:

i. *NP + Ø + Vt > PN + Ø + Vt*

ii. *NP/PN + Ø + Vt > NP/PN + Ø + Uner*

iii. *NP/PN + Ø + Ving > NP/PN + Ø + Ved*

4.3.3 Influence of Syntactic Environments on Ungrammatical Use of *BE*

In order to determine the extent of the influence of the syntactic environments on the ungrammatical use of *BE*, this sub-section re-examines the occurrences of overgeneration and omission in relation to the constituents occurring before and after *BE*.

4.3.3.1 Overgeneration of *BE*

Table 4.49 below summarises instances of overgeneration of finite *BE* in relation to the syntactic environments.

Table 4.49: Distribution of Overgeneration of *BE* According to Syntactic Environments in L1-Malay Data

Category	Item	Token	%
Subjects	Noun (NP)	151	45.76
	Pronouns (PN)	179	**54.25**
Form of Post-*BE* Verbs	Uninflected	206	**62.42**
	Inflected	124	37.58
Class of Post-*BE* Verbs	Transitive (Vt)	189	**57.27**
	Unergative (Uer)	91	27.58
	Unaccusative (Uac)	50	15.15
Intensifiers	Degree Adverbs	70	21.21
Modal Auxiliaries	Modals	23	6.97

4.3.3.1.1 Type of Subjects

Overgeneration of *BE* in the L1-Malay learner sub-corpus occur more often after PN subjects (54.25%) than NP subjects (45.76%). However, the difference in percentage is fairly small (8.5%), indicating that there is a general tendency for overgeneration of finite *BE* to occur after both types of subjects. This finding suggests that the type of subjects has no influence on the overgeneration of *BE*.

4.3.3.1.2 Form and Class of Post-BE Verbs

In terms of the forms of the post-*BE* verbs, finite *BE* is overgenerated more frequently with uninflected verbs (62.42%) than with inflected verbs (37.58%). Overgeneration according to Ionin and Wexler (2001) could be the result of learners' attempts to realise the inflection projection (IP) in their grammar. *BE* is believed to be inserted before a lexical verb as a mechanism for the learners to check the tense feature and to mark agreement (Ionin & Wexler, 2001; Lardiere, 1998). Ionin and Wexler (2001), stressed that learners who have not mastered the affixal inflection paradigm tend to resort to suppletive inflection (*BE*) to mark agreement and check the tense feature. They also added that the learners could be treating *BE* as a default for marking agreement, since they might be having difficulties accessing the affixal agreement. This can be traced by the behaviour of their overgeneration instances, which would often involve *BE* being overgenerated with uninflected lexical verbs as in *BE + V*.

In general, finite *BE* overgeneration constructions in the L1-Malay learner sub-corpus resemble the overgeneration found in the study conducted by Ionin and Wexler (2001), whereby *BE* is inserted before a lexical verb in the attempt to mark agreement and/or tense. Qualitative analysis of the overgeneration instances reveals that there appears to be almost no inconsistency in the subject and *BE* concord suggesting the possibility that learners are marking agreement with the use of suppletive inflection.

As for the class of post-*BE* verbs, overgeneration tends to occur more often with transitive verbs (57.27%) compared to unergative (27.58%) and unaccusative (15.15%) verbs. Previous studies done by Yip (1994), Balcom (1997), Oshita (2000), Ju (2000), Hirakawa (2006) and Park and Laskhmanan (2007) reported and discussed *BE* insertion before unaccusative verbs to produce *BE + Ven*, which the researchers termed as overpassivisation. The term overpassivisation is used to refer to this type of *BE* insertion as they resemble the English passives (Oshita, 2000). Oshita (2000) argued

that overpassivisation of unaccusative verbs is the result of learners' attempts to mark NP movement. In English, passive is marked by the movement of NP and the use of *BE + Ven* structure to do so. Unlike the overpassivisation construction in Yip (1994), Balcom (1997), Oshita (2000), Ju (2000), Hirakawa (2006) and Park and Laskhmanan (2007), the overgeneration constructions in the L1-Malay learner sub-corpus do not occur exclusively with unaccusative verbs, but they occur more frequently with transitive verbs. The difference perhaps lies in the underlying purpose of the overgeneration. The overpassivisation attested in Oshita (2000), Ju (2000), Hirakawa (2006) and Park and Laskhmanan (2007) is believed to be an overt evidence of learners attempt to assign a causal agent to the unaccassative verb or the failed attempt to passive the unaccusative verb. In contrast, *BE* overgeneration in the L1-Malay learner data is believed to derive from the learners' attempts to mark agreement, which is commonly realised in *BE +V* structure. The patterns of *BE* overgeneration in this study suggest that overgeneration is not sensitive to the class of post-*BE* verbs. The insertion of *BE* before a lexical verb is used as a mechanism to mark agreement, thus, overgeneration can occur with either transitive or intransitive verbs.

4.3.3.1.3 Presence of Modal Auxiliaries and Intensifiers

Overgeneration tends to occur more frequently without the presence of intensifiers and modal auxiliaries. The occurrences of finite *BE* overgeneration in these constituents are fairly low (28%). In general, overgenerations of *BE* in the L1-Malay learner data are not sensitive to the presence intensifiers and modal auxiliaries.

4.3.3.1.4 Summary of the Influence of Syntactic Environments on the Overgeneration of BE

The findings of this study indicate that *BE* overgeneration constructions occurring in the corpus data of L1-Malay ESL learners may be the outcome of a developmental aspect of language acquisition. It can be traced from the system underlying the patterns of

overgeneration, which are clearly made up of non-random constructions governed by very specific interlanguage grammar. As explained in the previous sub-section, *BE* is more frequently inserted before uninflected lexical verb in *BE + V* structure, where *BE* functions as an agreement marker. This overgeneration structure is consistently used by the learners with both transitive and intransitive verbs, showing evidence that the *BE* is inserted by the learners to project the English IP system, therefore, not the results of the influence of the syntactic environments.

4.3.3.2 Omission of *BE*

Table 4.50 summarises instances of omission of finite *BE* in relation to the syntactic environments.

Table 4.50: Distribution of Omission of *BE* According to Syntactic Environments in L1-Malay Data

		Copular		Auxiliary	
Category	**Item**	**Token**	**%**	**Token**	**%**
Subjects	Noun (NP)	48	59.26	71	65.14
	Pronouns (PN)	33	40.74	38	34.86
Subject Predicates	Noun Phrase (NP)	21	25.93	NA	NA
	Adjective Phrase (AP)	49	60.49	NA	NA
	Prepositional Phrase (PP)	6	7.41	NA	NA
	Wh-Clause (WhC)	5	6.17	NA	NA
Class of Post-*BE* Verbs	Transitive (Vt)	NA	NA	80	75.94
	Unergative (Uer)	NA	NA	28	23.85
	Unaccusative (Uac)	NA	NA	1	0.92
Intensifiers	Degree Adverbs	36	44.44	23	21.1
	Negation not	18	22.22	6	5.5
Auxiliaries	Modals	5	6.17	6	5.5

4.3.3.2.1 Type of Subjects

Consistent with the findings of Herat (2005) and Ellis (1988), L1-Malay learners in this study also tend to omit *BE* more often after NP subjects compared to after PN subjects. The trend is consistent for both copula and auxiliary *BE* omissions. There appears to be a strong relationship between the types of subjects and omissions of *BE*; NP subjects seem to exert a strong influence over omissions of *BE*.

According to Tode (2003), *pronoun + BE* sequenced are easier to memorise and supply as chunks. Specific pronoun would be sequenced with a specific morphological form of *BE,* for instance *it is..., they are...* or *you are...* and these combinations are very often reduced to contractions *it's, they're* and *you're,* which according to Wilson (2003) are acquired and supplied as chunks. The supply of *BE* after noun subjects is not automatic as nouns are not acquired as formulaic sequences such as *money is* or *computers are.* Determining the correct morphological form of *BE* after NP subjects is not a straight forward process, as the learners would have to first determine if the noun is singular or plural before they can supply the corresponding morphological form of *BE*. This makes supplying *BE* after NP subjects more challenging and may result in *BE* being dropped.

4.3.3.2.2 Subject Predicates

In this study, the influence of subject predicates only concerns copula *BE* constructions. Copula *BE* omissions are found to occur mostly before adjectival predicates (60.49%) and some before nominal predicates (25.93%). Interestingly, the same copula *BE* omission pattern is also found in the data of learners from other language backgrounds including Chinese (Lee & Huang, 2004), Sinhala (Herat, 2005), Russian (Unlu & Hatipoglu, 2012) and Arabic (Murad & Khalil, 2015). The fact that speakers from structurally different L1 backgrounds share similar *BE-adjective* constraint suggests that L2 learners in general share similar developmental pattern in the acquisition of the English IP system. This also means that the types of subject predicates to an extent influence the omissions of copula *BE* in this study.

4.3.3.2.3 Class of Post-BE Verbs

Another constituent that is analysed in relation to *BE* omissions is the class of post-*BE* verbs. This analysis only involves instances of auxiliary *BE* omission. As mentioned earlier, unaccusative verbs are believed to influence the erroneous insertion of *BE* (Ju, 2000; Oshita, 2000; Yip, 1994). In order to ascertain if the same influence exists,

omissions of *BE* are also analysed in relation to the post-*BE* verbs. The figures suggest no such influence as omissions of auxiliary *BE* involved mainly transitive verbs (75.94%) and very few involved unaccusative verbs (0.92%). The figures clearly show that class of post-*BE* verbs has no influence over auxiliary *BE* omissions in this study.

4.3.3.2.4 Presence of Intensifiers and Modal Auxiliaries

The other two constituents analysed with regard to omissions of *BE* are the presence of intensifiers and modal auxiliaries. Similar to the L1-Chinese learners in Lee and Huang (2004), L1-Malay learners are also found to omit copula *BE* in the structure *BE + very/not/so + adjective.* Approximately 44.44% of copula *BE* and 21.1% of auxiliary *BE* omissions occur in the presence of degree adverbs such as *always, so* or *very.* This finding indicates that degree adverbs have some influence over omissions of both copula *BE* and auxiliary *BE*. Nevertheless, omission instances in the presence of modal auxiliaries as those reported in Chan (2004) are very few in the L1-Malay data suggesting that this constituent has no influence over omissions.

4.3.3.2.5 Summary of the Influence of Syntactic Environments on Omission of BE

The findings suggest that omissions of *BE* are strongly triggered by the type of subjects. NP subjects are found to generate more instances of omission compared to PN subjects. According to Tode (2003) and Wilson (2003) this is directly associated to pronouns being acquired in formulaic sequences such as *they're* and *he's*, thus, would also be supplied as formulaic sequences without the need for the application of agreement rule. In contrast, the supply of the correct morphological form of *BE* preceding noun subjects would require application of the agreement rule (Tode, 2003). As a result, there is higher tendency for *BE* to be supplied correctly with pronoun subjects compared to with noun subjects (Tode, 2003).

The findings also suggest possible influence of subject predicatives on omissions of *BE*. *BE* is found to be absent more frequently in the *BE-adjective* structure compared to in

the *BE*-noun structure in the L1-Malay learner data. The same tendency was also reported in previous studies (Herat, 2005; Lee & Huang, 2004; Murad & Khalil, 2015; Unlu & Hatipoglu, 2012). This inclination seems to demonstrate a finer existence of developmental patterns in the acquisition of copula construction with *BE-noun* structure being perhaps more easily acquired compared to *BE-adjective* structure (Lee & Huang, 2004).

Finally, the supply of copula *BE* is also affected by the presence of intensifiers especially the use of degree adverbs. Copula *BE* is often omitted in the *NP + intensifier + AP* construction, where the adjective is modified by a degree adverb such as *so, always* and *very*, similar to the pattern of copula *BE* omissions recorded in L1-Chinese learner data in Lee and Huang (2004). The findings from the analysis of constituents before and after null *BE* suggest possible influence from the subjects, subject predicatives and intensifiers on the instances of *BE* omission in this study.

CHAPTER 5

RESULTS OF THE QUALITATIVE ANALYSIS

5.0 Introduction

L1-Malay ESL learner data were qualitatively analysed to find out how *BE* is actually used by the learners, which will provide the answers to the third and fourth research questions of the study. The analysis will highlight the patterns of the grammatical and ungrammatical uses of *BE* and the influence of the syntactic environments on the grammatical and ungrammatical uses of *BE.*

5.1 Grammatical Use of *BE*

This section presents and discusses the findings of the grammatical use of the major functions of finite and non-finite *BE*.

5.1.1 Grammatical Use of Finite *BE*

Finite *BE* mainly functions as either a copular, an auxiliary, a negation operator or an interrogative operator. In addition, *BE* is also used in the construction of existential *there* and *it*-cleft. This section presents and discusses in detail the qualitative findings of these functions.

5.1.1.1 Copula *BE*

According to Hinkel (2002) clauses with *BE* as the main verb are relatively simpler than those with verbs that have higher semantic and lexical content. Hinkel (2002) added that *BE* clauses produced by ESL learners also have characteristics of spoken rather than written discourse. She reported that L2 learners tended to construct copula *BE* structures in simple propositions made up mostly of a subject and a predicate and there was also

the inclination for *BE*-copula to be complemented by adjectivals as illustrated by the following sentences:

a. *The money making **is** important of course*.

b. *The contents in university **are** very difficult.*

Hinkel (2002, p. 114).

In contrast, corpus-based findings of Biber et al. (1999) reported that *BE*-copula constructions are more common in academic prose than in conversation. They are used mainly to express relationship between the subject of a clause and an attribute. They have several major functions, which change according to the types of complements preceding the copular (Biber et al., 1999), thus, can be part of structurally complex constructions. The following sub-sections present the patterns of *BE*-copula constructions in the L1-Malay learner sub-corpus.

5.1.1.1.1 The Patterns of Copula BE Constructions

The quantitative analysis of L1-Malay learner sub-corpus reveals more frequent occurrences of *BE*-copula constructions with nominal (*BE + NP*) and adjective predicates (*BE + AP*). Some of these constructions resemble the propositions characterised by Hinkel (2002) as simple and conversational. The simple *BE +AP* construction is used mainly to express specific evaluation (Biber et al., 1999) as exemplified in extracts (1) to (3) below:

(1). In our life, money ***is*** very important. E0024-05

(2). As we know, money ***is*** important to everyone. FP0054

(3). With imagination, the world ***is*** never dull. C00021

Meanwhile, the simple *BE + NP* construction is used to characterise the subject noun phrase (Biber et al., 1999) as exemplified by the extracts (4) to (5) below:

(4). Malaysia ***is*** a multiracial country consisting people of different races, custom and believes. FP0178

(5). The punishment ***is*** a minimum 15 to 30 years jail and minimum 10 strokes of the rotan. C0024

BE-copula constructions in the L1-Malay learner data, however, are not always expressed in simple propositions as exemplified in extracts (1) to (5) above. *BE* can be parts of syntactically complex structures. The subsequent sections discuss these complex *BE*-copula constructions further.

5.1.1.1.1.1 Copula *BE* in Complex Sentences

Complex sentences are formed with one main clause and one or more subordinate clauses that are connected together with subordinating conjunctions such as *although, because* and *before*. They can also be formed with two clauses that are connected by relative pronouns (e.g. *which, who, whose*) or relative adverbs (e.g. *where, when, why*) (Biber et al., 1999). The following sub-sections discuss the occurrences of *BE*-copula in the main and subordinate clauses of complex sentences.

5.1.1.1.1.1.1 Copula BE in Main Clauses

Firstly, the analysis of *BE*-copula in complex clauses focuses on the utilisation of the verb in the main clauses. The function of copula *BE* as a main verb is determined mainly by the predicative used to complement the verb. When followed by a noun phrase, copula *BE* is used to either characterise or identify an attribute as exemplified in extracts (6), (7) and (8) below. *BE* in extracts (6) and (7) is used to identify the subjects, while in extract (8) it is used to characterise the subject. The main clauses are indicated by square brackets [], while the subordinate clauses are underlined.

(6). [Prison ***is*** just **one of a number of sanctions** available to the courts to deal with those] who commit criminal offences. L0008

(7). [Arts and music ***are*** **two subjects]** that are believed *(to be)* responsible for developing imagination and creativity. FP0185

(8). [The children ***are*** **the most vulnerable of all**], as they are easily influenced and might carry out these undesired acts on themselves or with their friends. A0004

Another common subject predicative of *BE* is adjective phrase. In academic prose, larger range of predicative adjectives are used to complement copula *BE* and they would most likely be used to express specific evaluation. It is also common for the predicates to be complemented by another clause or prepositional phrase (Biber et al., 1999). As can be seen in extract (9), predicative adjective (*able*), which supplies the evaluation of the subject NP (*women and men*), is further complemented by several infinitive clauses (bold and italicised). Extract (10) is a sample of predicative adjective complemented by a prepositional phrase (*with ourselves*).

(9). [Women and men ***are*** **able** ***to choose to work, stay home,*** or ***to do both,*** and ***to help change society*****]** so that stay at home parents are not devalued or looked down upon. A0007

(10). The majority of people out there may tell us "money is everything" but if [all of us ***are*** **truly honest** ***with ourselves*****]** and earnestly seek the voice of conscience deep in us, we would discover that this statement cannot be true. A0018

5.1.1.1.1.1.2 Copula BE in Subordinate Clauses

The second part of the analysis focuses on the functions of *BE* in subordinate clauses. Subordinate clauses are dependent clauses that are normally embedded as part of another main clause. They can be signalled by a subordinator *that* or *wh*-word, by non-finite verb phrases introduced by infinitive, *-ing* participle or *–ed* participle and subordinating conjunctions such as *although*, *before* or *because* (Biber et al., 1999). According to Hinkel (2002) subordinate clauses are prevalent in academic texts and often used and regarded as the markers of textual and structural complexities. In writing, subordinate clauses are classified into three types; noun, adjective and adverb clauses and they are distinguished according to their textual and structural functions.

Noun clauses are marked by omitted and explicit *that*-clause or *wh*-clause and are most common in academic prose. They usually follow reporting verbs in summaries, restatements and citations (Leki, 1999; Swales & Feak, 2012). Other than that, they are also used to impart detachment and objectivity in text (Tadros, 1994), provide for extensive cohesive ties by recapitulating earlier information, predict development of discourse moves and delay proposition to the secondary clause position (Franchis, 1994). Extracts (11) and (12) below provide the samples of how copula *BE* is used as the copula of noun clauses functioning as the subject complements (Quirk et al., 1985). Extract (11) attempts to recapitulate earlier information, while extract (12) is used to delay proposition.

(11). Mostly the *kidnapers* (kidnappers) will want the parents to pay about thousand or even millions for the ransom. As we can see here, the similarity of these cases like snatching, robbery and kidnapping is that they ***are*** all about money. A0002

(12). In Malaysia, public interest in prisons ***is*** almost non-existent, probably because of the general view that the prison ***is*** the best place for criminals to go. L0038

Similar to adjectives, adjective clauses also function as post-positional noun modifiers (Hinkel, 2002). Extracts (13) is a sample of a copula *BE* in an adjective clause, which is post-modifying an object (*things*).

(13). Nowadays the criminal believes in **things** that ***are*** expensive even though they cannot afford it, and crime ***is*** the only means to achieve this aim. L0008

Adverb clauses have many contextual functions namely to express cause (*because, since, for*), concession (*although, though*), condition (*if, whether, unless*) and purpose (*so, so that*). Nevertheless, they are more common in speech than in writing (Biber, 1988). Nonetheless, they are occasionally used in academic writing as exemplified in extract (14), in which the adverb clause is used to express purpose.

(14). They will steal some money in order to have a 'high class' image so that they ***are*** able to be accepted in the social ring of high class and rich people. FP0055

5.1.1.1.1.1.3 Copula BE Used with Non-Finite BE

In many cases *BE* is used alongside non-finite *BE* (*be, been, being*) in a sentence. The non-finite forms can be used in passive voice (*modal + be + PP*), perfect passive (*have + been + PP*) or progressive passive (*be + being + PP*). They can either occur as independent clauses coordinated by conjunctions or as subordinate clauses normally serving as expansion to the main clause. Extracts (15), (16) and (17) are samples of compound sentences with each consisting of a copular construction joined by a coordinator (bold) to a clause containing non-finite *BE* construction (underlined).

(15). Violence ***is*** as bad as sexual explicity, **and** the truth ***should not be hidden*** from the society. A0001-05

(16). However in my point of view censorship in Malaysia ***is*** a little too hard **and** it ***had been said*** one of the toughest. A0008

(17). Men ***are*** no longer the leaders **but** ***are being led*** like little boys by unpleasant women in pants and short hair. A0020

Extracts (18), (19), (20) are the samples of complex clauses in which *BE* functions as copular (bold and italicised) of the main clauses that are modified by subordinate clauses having non-finite *BE* clauses.

(18). The answer ***is*** nothing but a huge bill incurred from the running of the centres that ***could be better used*** to feed the mouths of many starving children across the country. L0046

(19). Almost everyday there ***are*** people who ***had been robbed*** and cases like bank robbery that we can hear in the news. A0002

(20). Even though Internet censorship ***is being carried*** out these days, it ***is*** inevitable in eliminating everything that needs to be censored. A0004

5.1.1.2 Auxiliary *BE*

Another major function of *BE* is as an auxiliary in the progressive aspect (*BE + Ving*) or in the formation of passive voice (*BE + Ved*). The following sub-sections discuss how *BE* is realised for these functions.

5.1.1.2.1 Auxiliary BE in Passive Voice

Auxiliary *BE* in L1-Malay learner sub-corpus are used more frequently in the formation of passives. Passive voice constructed with auxiliary *BE* (*BE + Ved*) is common in the sub-corpus. According to Swales and Feak (2012), passives are traditionally more suitable than active voice in academic prose. In fact, they are often considered as requisite in written genres (Hinkel, 2002). Passives can occur with or without an agent. Agentless passives also known as short passives, occur without the agent being specified. Long passives, on the other hand contain a *by*-phrase in which the agent is specified (Biber et al., 1999). Passive constructions in L1-Malay learner sub-corpus are found to favour short passives than long passives.

5.1.1.2.1.1 Auxiliary *BE* in Short Passives

According to Biber et al. (1999), short passives, are about six times as frequent as long passives. They are very extensively used in academic prose, such as in academic research articles, in which the presence of human actor (agent) is not required or not important (Biber et al., 1999). The qualitative analysis of the short passives in the L1-Malay learner sub-corpus reveals that they occur mostly in complex clauses and some can be found in compound sentences. Extract (21) below is a sample of short passive in a compound sentence. The clause contains two compound sentences (separated by square brackets); $(21)^a$ contains two independent clauses that are both passives, while $(21)^b$ is a combination of an active clause (*where is the justice to this …*) and a passive clause (*this movie was banned …*).

(21). [For example last year, the comic character movie Daredevil ***was banned*** but the comic ***was sold*** like normal,][a] [where ***is*** the justice to this matter and this movie ***was banned*** at last minute after the promotion,][b] these really make a big lost for our cinema industry. A0008

Nevertheless, the data reveal that it is more common for the passive clauses to also be the dependent clauses of complex sentences. As exemplified in extracts (22) and (23), the nominal clauses (underlined and italicised) are in the passive voice.

(22). The expectation is *that the university leaders **are drawn** from the best brains in society* and they can play the integrative multiple roles of being the perpetual learner, researcher and teacher. A0011

(23). There are some *who **are** so **caught up** with earning money* that they neglect spending time with their families and loved ones. A0018

Occasionally, the short passives can be found in the main clauses as well as in the subordinating clauses as exemplified in extract (24) below:

(24). In university, courses ***are conducted*** in such a way *where 100 percent marks **are not fully based** on examination papers*. B0023

5.1.1.2.1.2 Auxiliary *BE* in Long Passives

According to Biber et al. (1999), the use of long passives is motivated by three interconnected principles. The first is information-flow principle or the preference to present new information at the end of a clause, after already-shared information. The second is end-weight principle, which means that the heavier or lengthier element is placed at the end of a clause. The third principle states that long passive is chosen to place initial emphasis on the theme of a discourse. All three principles tend to support one another (Biber et al., 1999).

These principles, however, do not seem to be consistently applied in the long passives constructed by L1-Malay learners in MACLE. The information-flow principle is observed to be upheld in extracts (25), (27) and (28), while the rest of the extracts do

not seem to be introducing any new information in the *by*-phrase (underlined). For instance in extracts (26) and (27) the main verbs provide very clear indication of the agents to come; common sense would inform readers that the verb *taught* would be followed by a teaching figure and similarly the verb *conducted* would summon for an authoritative body like the university. Nevertheless, in extract (28) one of the agents, *jealousy* has already been introduced in the previous sentence, while the other agent *mentally-ill people* is new information. The end-weight principle, which refers to the lengthier elements of the clause being placed at the end, also appears to be violated in all extracts except in (29). As for the third principle, it appears to be adhered to in all except extracts (25) and (28). Nevertheless, infringement of these principles does not affect the grammaticality of the passive constructions (Biber et al., 1999).

It is observed that learners' tendency to opt for long passives is also lexically constrained, for instance verbs like *grip* (25), *govern* (28) and *do, cause* (29) require the agent to be present unlike for instance *teach* (26) and *conduct* (27).

(25). We ***are gripped*** by fear and the criminals have a relatively easy time taking advantage of the weaknesses of the system. A0006-05

(26). In another word, not only we learn about what ***are written*** in the text books or what ***are taught*** by the lecturers, but we also need to grab this opportunity to learn things beyond that in order to survive out there. B0023

(27). There are many seminar and workshops that ***are conducted*** by university for their student. B0038-05

(28). Being an Islamic country, we ***are governed*** by the basic laws of Islam,… B0197

(29). According to the Christian and Muslim faith, the first evil deed ***was done*** by one of Adam's son who killed brother out of jealousy as he wished to marry the bride whom his father had chosen for his brother. As for today, when we read the newspapers, many evil deeds and murders committed ***were mostly caused*** by jealousy or by mentally-ill people. A0016

Long passives in the L1-Malay learner sub-corpus rarely occur in simple sentences. Nevertheless, they can still be found occasionally in the data as exemplified in extract (30) below. The inclusion of the agent provides more emphasis on the party (*their governments*) responsible for giving the permission to sell weapons to the public.

(30). The weapon industries ***are even permitted*** by their governments to sell weapons to public. B0059-05

Long passives in compound sentences are also not very common in the data. As shown in extract (31) below, the independent clause in which the long passive transpires (marked []) is coordinated by *and* to another independent clause (underlined).

(31). [We ***are gripped*** by fear] ***and*** the criminals have a relatively easy time taking advantage of the weaknesses of the system. A0006-05

The qualitative analysis reveals that most of the long passives can be found in complex sentences. Learners' general preference for complex sentences results in comparatively common occurrences of long passives in this construction. In addition, most of the long passives in the L1-Malay data are heavily dependent on the significance of the agents. As can be observed in extracts (32) to (34) below, the agents (bold and underlined) in the by-phrase are complemented by nominal dependent clauses (italicised), which provide further description/information of the agents. The description/information places important focus on the agents and draws the readers' attention to them. Most of the long passives found in the learner data are constructed in this manner.

(32). According to the Christian and Muslim faith, the first evil deed **was done** by one of **Adam's son** *who killed brother out of jealousy as he wished to marry the bride whom his father had chosen for his brother*. A0016

(33). Crimes related to money **were mostly done** by the **illegal money lenders** *who had no choice but had to resort to violence in order to collect the sum owed to them*. A0016

(34). The destruction of the family **was planned** carefully by **those** *who see the role of husband as an obstacle to state control*, or the Government husband mentality. A0020

5.1.1.2.2 Auxiliary BE in Progressive Aspect

Progressive aspect is mostly associated with spoken and informal register (Biber et al., 1999) and often employed in personal and/or expressive narratives (Hinkel, 2002). Despite this tendency, progressives are occasionally found in academic prose to indicate action in progress at a given time either in the present or past (Quirk et al., 1985) and they are often marked by auxiliary *BE + Ving.*

Present progressive (*am/is/are + Ving*) is used to describe events currently in progress or events that would take place in the future, while past progressive (*was/were + Ving*) is used to describe events that were in progress in a certain duration of time in the past (Biber et al., 1999). In the L1-Malay learner sub-corpus, present progressive constructions occur more frequently than past progressive constructions. This could be due to the topics learners were required to write, which deal mostly with current issues. This leaves the learners with little opportunity to employ the past tense, compared to when writing a historical or a narrative essay.

Typically, the progressive aspect is employed to express one or more events in progress in the present time as exemplified in extracts (35) and (36) below. Nevertheless, learners also employed progressive aspect to express continuous action that may have begun in an unspecified time in the past and still in progress at the present time as exemplified by extract (37). It is understood that the act of censoring has been in progress and still continues to occur in the present time. The employment of progressive aspect gives the reader a sense of presence, and places the reader in the scene depicted by the writer. It is also as though the writer is interacting with the reader. This effect will not have been possible with the use of for instance simple present tense.

> (35). We should not simply assume that if some people tell us that the sun ***is shining***, they are right. Even when it***'s raining***…However, by censoring these publications, what we ***are doing*** is actually clamping down discussions on these issues. L0010

(36). One ***trying*** to make money, while the other one ***is spending*** it. Money has been seen as the reason why our family institution ***is rocking, breaking*** apart and collapse. B0215

(37). Censorship ***is taking*** away the rights of citizens; it ***is protecting*** the rights of people who do not wish to be exposed to certain things. C0018

As for the types of sentences in which progressive auxiliary *BE* occur, the data reveal that they are occasionally found in simple sentences as exemplified in extract (38).

(38). ***Are*** they ***still investigating***? A0006-05

Progressives are also rarely available in compound sentences. Extracts (39) and (40) are samples of progressive constructions occurring in compound sentences.

(39). What they***'re planning*** to do is a crime, but they still do it because they know the laws. C0001

(40). To me, happiness ***is spending*** quality time with loved ones and ***doing*** something that you like and the most important thing ***is honouring*** and ***obeying*** God. A0018

Comparatively, progressives are more common in complex constructions. They can either be positioned in the main clause, in the subordinate clause or in both. In extract (41) below, progressive is employed in the dependent clause (underlined), while extract (42) is a sample of progressive aspect in the main clause

(41). You stand in the midst of the greatest achievements of the greatest productive civilization and you wonder <u>why it***'s crumbling*** around you while you***'re damning*** its life-blood money</u>. C0030-05

(42). Employers nowadays ***are no more seeking*** for graduates who pass with flying colours. B0090

5.1.1.3 *BE* in Existential *there* Constructions

Existential *there* is used as a device to state the existence of non-existence of something (Biber et al., 1999). Its main function is to introduce new information, which is presented in the notional subject or the indefinite noun phrase proceeding *BE*.

Existential clauses sometimes have definite noun phrases or proper nouns. Typically, existential *there* clause is formed in the following structure:

there + *BE* + indefinite noun phrase (+ place or time preposition adverbial)

There's a bear sitting in the corner.

Biber et al. (2002, p. 412)

The notional subject can be structurally simple or complex. A simple notional subject is normally not modified by any post-modifying clause, while a complex one can be expanded with post-modification clauses or with adverbials (Biber et al., 1999).

Simple existential clauses as in extracts (43) and (44) below are very rare in the L1-Malay learner sub-corpus, while existential clauses expanded with adverbials are not found in the in the sub-corpus. Adverbials in existential clauses provide the information of when and where something exists and they are often placed at the end of the clause (e.g. *There are no trains **on Sundays***) (Biber et al., 2002, p. 415). The majority of the existential clauses found in the L1-Malay learner data are expanded with post-modifying nominal clauses such as *that*-clause (italicised and underlined) as in extracts (45) and (46), infinitive clauses as in extract (47) and relative clauses as in extract (48).

(43). If we think deeply, there ***is*** a connection! B0213

(44). There ***are*** different types of censorship; among them are internet, music, television, film, movie and radio censorship. A0004

(45). There ***are*** some ways *that can be recommended to overcome this problem* which make the non academic subjects or activities compulsory to all students. B0109-05

(46). So there ***is*** no surprise *that many children of well-known people have been accused by many offences such as drug abusing, raping, slaughter, corruption, cheating and murder.* B0117

(47). More prisons are being built around the world because [there ***is*** not enough room *to hold inmates*]. K0015

(48). …now there **is** telephone *which can use to communicate with other people in different places*. F0114

The majority of *BE* in the existential clauses in the L1-Malay learner data functions as copular as exemplified in extracts (45) to (48) above. Nevertheless, they can also be preceded by modal auxiliaries (*will, may*) or auxiliary *have* (Biber et al. 1999) as exemplified in extracts (49) to (51) below:

(49). If today, people stops dreaming and imaganing (*imagining*) things, *there* ***will be*** *no improvement in life*. FP0061

(50). While *there* ***may not be*** *as many resources available for victims of property crimes as other victims*, these individuals do have many rights and services available to them. K0015

(51). *There* ***have also been*** *rather lengthy episodes* when the opposite seemed true, when economic disruption apparently stemmed not from too little money, but from too much of it. I0011

5.1.1.4 *BE* as Negative Operator

BE also functions as a negative operator when used with participle *not* as in *is not, are not, was not* and *were not*, and they can be contracted to *isn't, aren't, wasn't* and *weren't* respectively. Generally, negative clauses are not common in written register (Biber et al., 1999), but are still used occasionally to negate sentences. The analysis focuses on *BE* as the negative operators in copular constructions, progressive aspect and passive voice.

5.1.1.4.1 Negation in Copula BE Constructions

Negative *BE*-copula constructions occur most commonly in complex sentences and occasionally in simple and compound constructions. Extract (52) below is a sample of negation in a simple sentence, while extract (53) is a sample of negation in a compound sentence.

(52). Depending on government agencies alone in crime *preventation* (prevention) ***is not*** enough. I0005

(53). To love a thing ***is not*** such a big problem because the love of money and money itself have a clear *dinstiction* (distinction). K0042

As for negation in complex constructions, it can occur in the main clause (54) or in the subordinate clause (55).

(54). [Sex, violence, abusive words and vulgarity ***is*** just ***not*** a norm in our culture], though in closed doors, it may happen, but we cannot allow this in our highly moral led society. B0197

(55). She gets into the car and drives to the nearest police station, only to be told [that it ***is not*** the right station to make the report]. A0006-05

Copula *BE* with negation can occur with a variety of complements, which include phrasal complements (nominal, adjectival and prepositional predicates), complement clauses (*that*-clause, *wh*-clause) or non-finite clause (*to*-infinitive clause). Extract (52) is a sample of negative copula *BE* complemented by an adjective predicate, while extracts (53), (54) and (55) are samples of *BE* complemented by nominal predicates. In general, nominal predicates are the most common complement proceeding copula *BE* in negative constructions and these noun phrases are often modified by post-modifiers such as a prepositional phrase as in extract (54) and *to*-infinitive clause as in extract (55).

5.1.1.4.2 Negation in Progressive Aspect

Negation in progressive aspect is not widespread, but can still be found in the data. The data reveal very limited use of negation in simple sentences, but it can be found in compound and complex sentences as exemplified in extracts (56) and (57) respectively.

(56). Money ***is not only driving*** the money needed people crazy but also the rich people. B0117

(57). Men who have no courage, pride of self-esteem, men who have no moral sense of their right to their money and ***are not willing*** to defend it as they defend their life, men who apologize for being rich will not remain rich for long. C0030-05

5.1.1.4.3 Negation in Passive Voice

In general, there are very limited occurrences of negation in the passive voice. Negation tends to also favour short passives compared to long passives. In fact

negation in long passives is almost non-existent in the data. Extracts (58) and (59) below are samples of negation in short passives.

(58). They ***are not even required*** to practice the skills they have been taught. E0025-05

(59). Students ***are not bothered*** about the cleanliness just as they ***are not bothered*** about being what they are suppose to be after graduated. FP0124

5.1.1.5 *BE* in *It*-Cleft Constructions

It-cleft consists of the pronoun *it*, a form of *BE*, a focused element (a noun phrase, a prepositional phrase, an adverb phrase or an adverbial clause), and a relative dependent clause introduced by *that*, *who/which* or zero (Biber et al., 1999). *It*-cleft constructions in the L1-Malay learner data occur in two major patterns *BE + copula + adjective* as exemplified in extracts (60) to (62) and *BE + copula + noun* as in extracts (63) to (66).

(60). Although **it is clear** [that the censorship is playing its role in sculpting a better society amongst us], some do not seem to agree. A0004

(61). **It *is* most impressive** [*that feminism has landed many women in schools, colleges and universities*]. FP0149

(62). **It *is* undeniable** [*that our society is now dominated by science, technology and industrialization*] and these are the driving force that governs our society to its present glory. L0031

(63). **It *is* the love of money** [*that can lead to evil*]. K0014

(64). **It *is not* money** [*that drives men to evil*.] **It *is* their love** for it. L0076

(65). **It is the hope of the society** [that the offenders will change their attitudes and respect laws and regulations existing in society]. L0049

(66). In the fast changing world of technology, **it is important to realise** [that there is a very high possibility that they may end up in areas they are not trained for]. S0040-05

It-clefts by NNS learners according to Hinkel (2003) were largely identified in *BE + copula + adjective* and relatively simpler than native speakers' prose, which was noted to contain *it*-clefts with "greater verbs and other attendant elements" (p. 291). As

attested by Hinkel (2003), the L1-Malay learners tend to produce relatively simple *it*-cleft constructions as exemplified in extracts (63) and (64). Nevertheless, there are also comparatively more complex *it*-cleft constructions available in the data as exemplified in extracts (60), (61), (62), (65) and (66). Extract (66) for instance, contains a wide range of attendant elements, which include a prepositional phrase (*In the fast …*), a complex adjective phrase which is modified by an infinitive clause (*important to realise..*) and subsequent relative clause (*that there is a very high…*).

Even though *it*-cleft constructions in the L1-Malay learner sub-corpus are not as widespread as other copular constructions, they are used consistently among more proficient learners. Considering that *it*-clefts are advanced syntactic construction (Hinkel, 2003) and when employed in scientific and academic prose marks the text as having relatively formal register (Scollon, 1994), the employment of the structure in the L1-Malay learner essays is a clear evidence that the learners are capable of syntactically complex *BE* construction.

5.1.1.6 *BE* as Interrogative Operator

Even though *BE* can used to form *wh*-questions, yes/no and tag questions, interrogatives in the L1-Malay learner data consist mainly of yes/no questions which have VS word order with *BE* as the operator (Biber et al., 1999) as exemplified in extracts (67) to (72). According to Biber et al. (1999), interrogatives are less common in writing and whenever they are used in the written register, they normally have rhetorical purposes (Biber et al. 1999). Interrogatives formed with *BE* are also found to be rare in the L1-Malay learner data. They are most commonly employed to engage the readers and involve them in the discussion or argument. Extracts (67) to (72) below are some interrogatives extracted from the L1-Malay learner data.

(67). Many factors can influence a person to commit a crime, but ***is*** *there* ***a common trait*** [*that leads people down the road to actually committing a crime*]. K0054

(68). ***Is*** *it* ***true*** [*that most university degrees are theoretical and do not prepare student for the real world*]? B0080

(69). Ask yourself, ***are*** *you man enough and willing to fight for your country*? Sacrifice your life in war? K0036

(70). Money is made possible only by the men who produce it. ***Is*** *this* [*what you consider evil*]? C0020

L1-Malay learners also tend to construct syntactically complex interrogative constructions as exemplified in extracts (67) and (68) above. Extract (68) for instance consists of an adjectival complement (*true*), which is post-modified by a relative clause (*that most university degree*…). Relative clause post-modification (*that leads people down the road*...) is also used in extract (67). Nevertheless, learners are also found to use simple interrogatives as exemplified in extracts (71) and (72) below:

(71). ***Is*** money bad? A0005

(72). But ***is*** it true? B0133

In general, interrogatives in the L1-Malay data are used for rhetorical purposes as attested by Biber et al. (1999). They are employed for three major purposes; to enforce the main point as in extract (67), to direct readers' attention to the essay topic as in extracts (68) and (69) and to form persuasion as in extract (70).

5.1.1.7 Summary of Grammatical Use of Finite *BE*

The findings from the qualitative analysis of *BE*-copula reveal the L1-Malay learners exhibit a considerably high level of competency in its constructions and functions. *BE*-copula constructions in the L1-Malay learners' essays are used according to the structures and functions discussed in Biber et al. (1999). Contrary to Hinkel (2002; 2003), who reported repetitive and overly simplistic constructions of *BE*-copula among L2 learners, L1-Malay learners use copula *BE* in complex sentences expressing

complex ideas. In fact, complex constructions are more common in the L1-Malay learner data compared to simple and compound constructions. This tendency surfaces from the need to express complex ideas/arguments in which case the main clause are often extended by one or more subordinate clauses. According to Hinkel (2002) subordinate clauses are prevalent in academic texts and often used and regarded as the markers of textual and structural complexities.

Auxiliary *BE* is more frequently found in the formation of passives than in progressive aspect. Passives, according to Swales and Feak (2012) are traditionally more suitable than active voice in academic prose and they are often considered as requisite in the written genres. According to Friginal, Li and Weigle (2014) higher distribution of passives in learners' writing is a good indication of quality writing. The researchers found that passive structures were preferred by both highly rated NS and NNS writers. In the L1-Malay learner sub-corpus, agentless passives are observed to be more preferred than long passives. Similar to *BE*-copula constructions, these passives appear to be more prevalent in complex sentences than in simple and compound sentences.

Due to the nature of progressive aspect which is mostly associated with spoken and informal register (Biber et al., 1999), they are generally scarce in the L1-Malay data. Nevertheless, some can still be found in the learners' essays and they are more common in complex constructions than in simple and compound constructions.

Other than its major functions; copular and auxiliary, finite *BE* forms in the L1-Malay learner data are also used as negative and interrogative operators and used in the construction of existential *there* and *it*-clefts. Existential *there* constructions are found to be comparatively more common in the data compared to negations, interrogatives and *it*-clefts. In general, learners are observed to exhibit considerable competency in all four constructions.

5.1.2 Grammatical Use of Non-Finite *BE*

Non-finite *BE* forms (*be, been, being*) are used by the L1-Malay learners mainly in the construction of passives. Passive constructions, according to Biber et al. (1999) are most common in academic prose and according to Hinkel (2002) is also a prominent feature in composition writing. Passive voice is often used to project the writer's objectivity in a style that is interpersonal, indirect and detachable. This section discusses and presents the data of non-finite *BE* in the L1-Malay learner data.

5.1.2.1 Infinitive *be*

5.1.2.1.1 Infinitive be in Passive Voice

Infinitive *be* is found to be the most common form used by the learners besides the finite *is* and *are*. The data reveal that the form is most commonly paired with modal verb *should* and *could* to produce passive construction *modal + be + Ved*. The construction is observed to perform two major functions; making suggestions or recommendations and expressing possibilities.

Extracts (73) to (75) are samples of suggestion making realised by *modal + be + Ved* construction. Learners' preference to put forward their suggestions in the passive voice is determined by how important is the agent and the necessity to include the agent in the construction. As shown in extracts (73) to (75) the need to include the agents does not arise, instead the focus and importance is placed on the action to be taken and the object. In addition, the use of the modal auxiliaries (*should*) transmits a sense of obligation to the object in executing the suggestions. Obligation modals, namely *must*, *should* and *have to*, according to Biber et al. (1999) are more common in academic prose and in this case the learners' preference to *should* is probably due to it being a less threatening modal in expressing obligation than *must* and *have to.*

(73). Therefore, in deciding the appropriate sentence, a court ***should always be guided*** by certain considerations such as public interest to curb the increasing of the statistic of offences…. In conclusion, the imprisonment system ***should not be abolished*** and ***replaced*** by the rehabilitation system. L0005

(74). Prison ***should be addressed*** through training and rehabilitation programs which seek to reduce the chances of released prisoners offending again. L0008

(75). Abortion clinics ***should be taxed***, to help provide for stay at home parents, and to help fund the unwed mother's home and adoption centers. A0007

The same construction is also used in expressing possibility as exemplified in extracts (76) and (77). Modal verb *could*, which is most common in the passive voice (Biber et al., 1999), is employed for this purpose. According to Biber et al. (1999), *could*, *may* and *might* are used exclusively to mark logical possibility. Nevertheless, there are very few instances of *may* and *might* being employed for this purpose in the L1-Malay learner sub-corpus. The following extracts provide samples of *could* used in expressing possibility.

(76). I believed that from the rehabilitation, the prison overcrowding or the quantity of the prisoners ***could be reduced***. L0008

(77). A good part of the annual criminal statistics ***could be prevented*** if social problems relating to unemployment, poverty, lack of education, proper housing, and so forth, are addressed by the government. B0126-05

5.1.2.1.2 Infinitive be in Active Voice

Infinitive *be* also occur in *modal + be* construction, which is mainly employed in suggestion making and expression of obligation. In this case, the use of modal auxiliary tones down the argument and is observed as a strategy undertaken by the learners to exercise caution in their arguments, thus, not appearing arrogant or too assertive. This is an important aspect of academic writing because a writer's aim is not only to persuade, but also to build positive writer-reader relationship, which can be achieved through the employment of hedging devices such as modal auxiliaries. Instead of using

definitive expression such as "*the solution is to address*" (78) or more assertive modal auxiliary such as *must* in the expression of obligation, learners have opted for less assertive modals, namely *would* and *should* in advancing their arguments as exemplified in extracts (78) to (80) below:

(78). The best solution ***would be*** one that would address the fine balance between the prisoner's individual rights and the interest of the society. L0010

(79). Therefore, the university degrees ***should be*** more on practical rather than theoretical. B0025-05

(80). The way teachers, lecturer or tutor present their lecture ***should be*** interesting that makes student involve with the topic. B0088

5.1.2.2 Non-Finite *been*

The non-finite *been*, which can only occur after auxiliary *have*, performs 3 major functions; present/past perfect tense (*have + been*), perfect progressive (*have + been + Ving)* and perfect passive (*have + been +PP*) (Biber et al., 1999). Textual analysis of *been* in the L1-Malay learner sub-corpus reveals that the form is most frequently used in the formation of perfect passive. According to Biber et al. (1999), perfect aspect and passive voice are both common in academic prose and news. Canonically, perfect passive is used to show past time with present relevance as exemplified in extracts (81) to (84).

(81). Imprisonment as a way to deal with criminals ***has been established*** since civilisation started. It ***has been proved*** to be the most humane and at the same time effective method to keep society safe from crime doers. L0025

(82). Always ***has been observed*** that the purpose of imprisonment is to incarcerate individuals who need a "timeout" from the society. L0049

(83). It is not absurd to say that money hold the power against those who ***have been slaved*** by it. S0045-05

(84). Music CDs with vulgar lyrics that ***have already been*** censored are labeled on the cover as a note of advisory for its explicit contents. A0004

Present perfect tense involving *BE* (*has/have been*) is also available in the learner data, but its occurrences are relatively low compared to perfect passive. Swales and Feak (2012) noted that the use of perfect aspect with present tense (present perfect) is conventionalised in academic writing and commonly employed to mark continuing actions in research, discussion or project. Present perfect realised in *has/have + been* often has copular function (Biber et al., 1999). Present perfect constructions in the L1-Malay learner sub-corpus are mainly employed as copular to indicate past action that started in the past and continues in the present as exemplified in extracts (85) to (87).

(85). Whilst attributes like creativity, enthusiasm and willingness to learn new skills are always quoted, many emphasize that the quality of local graduates ***has been*** on decline and nowhere near what is needed for the K-economy (knowledge-based economy). S0040-05

(86). Woman ***has been*** a companion of man for a very long time. A0020

(87). For the human male is and always ***has been*** a hunter, he started out hunting large animals, other men and women, and went on to hunt money, other men and women. K0001

Past perfect *had + been* is found to be extremely rare in the L1-Malay learner sub-corpus. In general, past perfect is commonly employed to indicate "past-in-the-past" (Quirk et al., 1985, p. 195) and they occur most often in historical narratives (Hinkel, 2002), but not often in compositions and academic texts.

5.1.2.3 Non-Finite *being*

The form *being* in the L1-Malay learner sub-corpus is mainly used in progressive passive (*BE + being + PP*) and according to Biber et al. (1999) progressive passive is comparatively rare, but it can be found occasionally in academic prose. Similarly, L1-progressive passive are very limited in the L1-Malay learner sub-corpus. Progressive passive is commonly used to express actions that are in progress or incomplete in the present, past and near future. It is interesting to discover that even though the form is

not widely used, whenever it is used, it is used correctly. Extracts (88) and (89) below are two samples of progressive passive found in the L1-Malay learner data.

(88). We would like to know if enough measures ***are being made*** to improve the current situation. A0006-05

(89). …when you see corruption ***are being rewarded*** and honestly becoming a self-sacrifice, you may know that your society is doomed. C0030-05

5.1.2.4 Summary of Grammatical Use of Non-Finite *BE*

The analysis non-finite *BE* forms in contexts reveals a very consistent pattern, which is all three forms are predominantly used in the formation of passives. Passive voice is especially common in academic prose and compositions (Biber et al., 1999: Hinkel, 2002), therefore, it is not surprising to discover that learners have exhausted all three non-finite *BE* forms for this purpose.

It is also important to highlight that modal auxiliaries are commonly employed as a hedging device in infinitive *be* constructions. In academic writing, it is common for writers to employ hedging devices, which include modal auxiliaries such as *could*, *would*, *should*. They are mainly used to soften arguments (Hyland, 2005; Hyland & Tse, 2004). The employment of these devices is not only welcomed but also encouraged as they help writers to adhere to the convention of academic writing. The use of hedging has become conventionalised and is a common feature of academic writing and writers conforming to this convention as attested in previous studies were better rewarded (Dana, 2008).

5.1.3 Overall Summary of the Grammatical Use of *BE*

The qualitative analysis of the grammatical use of *BE* permits for further investigation on learners' proficiency to be conducted. Generally, learners who employ syntactically more complex *BE* constructions are observed to score Band 4 and higher in the

Malaysian University English Test (MUET)[1]. Both finite and non-finite *BE* forms are used in a greater range of syntactic constructions as exemplified in extract (90), which was extracted from an essay by a Band 5 learner.

(90). Crime happens for many reasons which for those who ***are*** in lack, struggling materially, crime tends to happens out of economic desperation and its consequences. A good part of the annual criminal statistics ***could be prevented*** if social problems relating to unemployment, poverty, lack of education, proper housing, and so forth, ***are solved*** by the government. Sociological studies and statistics have demonstrated a clear connection between crime, violence and the unequal distribution of resources in modern societies. C0009 B5

The excerpt above contains three *BE* constructions; copular, future passive and present perfect, which are woven with other verbs (underlined) with higher semantic and lexical content than *BE* (i.e. *happen, tend, demonstrate*). This demonstrates that learners especially those with higher proficiency in English employ a wider range of verbs, which *BE* is a part of. In addition, as can be seen from extract (90), *BE* is employed as a copular only once in the entire paragraph, which is contrary to the findings of Hinkel (2003) who reported an extensive use of simplistic and conversational style *BE*-copula constructions in the L2 learners' writings she analysed.

Extract (91) below provides another sample of the writing of a more proficient learner, in which *BE* is used effectively in constructing a rhetorical question (*Is that evil?*), which introduces the reader to the arguments forwarded by the writer. In addition, *BE* is only employed twice in the paragraph; once as a copular (*Is that evil?*), and in the formation of simple future (*would be*), which clearly shows that the learner is not relying solely on *BE* constructions in advancing his arguments.

[1] The **Malaysian University English Test (MUET)** is a test of English language proficiency for university admissions. It is a prerequisite in applying for admissions into all public universities and colleges in Malaysia. The scores are graded in 6 bands with Band 6 the highest and Band 1 the lowest. Band 6: Very good user; Band 5: Good user; Band 4:Competent user; Band 3: Modest user; Band 2: Limited user; Band 1: Very limited user.

(91). A man strive for excellence in his life; excellence in the business or corporate world. He may seek to better himself for the sole objective of making as much money as possible. ***Is*** that evil? It depends. He may, once he reaches the pinnacle of his wealth, use that wealth to commit unspeakable atrocities. Even in the pursuit of that wealth, he may resort to questionable deeds. Then, that ***would be*** an evil thing which he has done. What if, however, we again take the same man who ventures on that same path to riches. Yet this time, once he reaches the heights that he has aimed for, he turns his mind to deeds charitable and philanthropic. The root of his drive: money. The results? L0076 B6

Similarly, non-finite *BE* forms are also employed by proficient learners with some restraint as can be seen in extract (91) and (92). Infinitive *be* is used only once in the formation of simple future in extract (91) above and three times in the constructions of passive voice in extract (92) below:

(92). In conclusion, the imprisonment system ***should not be abolished*** and ***replaced*** by the rehabilitation system. Both should go together to ensure the efficiency of the sentence. Therefore, the government should take certain steps to upgrade these two systems such as steps ***should be taken*** to ease overcrowding throughout the prison system. In addition, because prisoners' ability to find work upon release into the community ***is*** an important determinant of their likelihood to commit future crimes, the authority should take greater care to provide work opportunities that help prisoners gain marketable skills. For example, in order to provide prisoners with a much needed outlet for self-expression, the authority should offer a program of arts and crafts in the prisons. By doing these, it will help the prisoner to face the new world and will not do the same offence. L0005 B4.

Analysis of the prose written by learners with Band 3 score in MUET (modest users of English), reveal slightly more frequent employment of *BE* compared to learners scoring Band 4 and higher. Nonetheless, the *BE* constructions do not dominate the learners' prose as shown in extracts (93) and (94) below. Extract (93) contains altogether nine verbs, seven of which are *BE*, while extract (94) consists of twelve verb phrases with only four *BE*.

(93). Censorship ***is not taking*** away the rights of citizens; it ***is protecting*** the rights of people who do not wish to be exposed to certain things. It ***is*** also a great tool in

preserving morals and social order. Violence in things such as movies and pornography ***are*** obvious to encourage criminal, or immoral behavior. Restricting such materials to certain times and places may keep them being viewed as taboo, and not allow them to become the norms of society. When these grounds ***are considered,*** we can see that censorship ***is*** a beneficial tool and ***must be applied*** in order to keep society at a safe, respectable, and just level. C0018 B3

(94). In things like movies nowadays, we can see a real merge and collaboration between technology and imagination. Movies ***are*** no longer just a movie like the olden days of 'Breakfast at Tiffany's' or movies by Humphrey Bogart. Now we see movies that ***are*** different from the ordinary like 'The Matrix', 'Star Wars Episode 1', and so on and so forth. Those movies show how technology ***is*** used to realise their imagination on screen to the people. For example, 'Star Wars' ***was constricted*** by limitation when they wanted to do Jaaba the Hut because moving the thing around like a person walking or something even close ***was*** impossible. Nevertheless, the moviemakers recently edited the movie and added Jaaba the Hut walking round Hans Solo. This shows our achievement in making it happen. E0001 B3

In contrast to the findings of Hinkel (2003), who reported heavy reliance of *BE*-copula constructions often complemented with adjectival predicates, which Hinkel concluded rendered L2 learner prose conversational and lacking elegance, proficient L1-Malay learners in this study exhibit the ability to use *BE* in syntactically complex constructions and show adequate restrain in the employment of the *BE* resulting in more varied and elegant use of the language as proved by extracts (90) to (94) above.

5.2 Ungrammatical Use of *BE*

Four major types of ungrammatical use of both copula *BE* and auxiliary *BE* have been recorded in the language data of the Malaysian ESL learners, namely omission, overgeneration, agreement and tense errors (Maros et al., 2007; Wee, 2009; Wee, Sim & Kamaruzam, 2010; Nor Hashimah, 2008; Siti Hamin & Mohd Mustafa, 2010). The quantitative analysis of the ungrammatical use of *BE* in the L1-Malay learner data in the MACLE reveals that only three types of errors occur most persistently, namely

omission, overgeneration and agreement. The following sub-sections present and discuss the findings of the qualitative analyses of two of these ungrammatical use of *BE* i.e. omission and overgeneration.

The analyses are divided into three major activities; the first focuses on examining the patterns of the errors. The second analysis involves investigating the errors in relation to the syntactic environments; examining the possible influence of the environments on the occurrence of errors. The final analysis involves examining the errors in relation to syntactic complexity or discovering the types of clauses (simple, compound or complex) in which the errors are more persistent.

5.2.1 Overgeneration of *BE*

As explained in previous section the analysis of overgeneration of *BE* in the study focuses mainly of the patterns that they take, the syntactic environment in which *BE* is overgenerated and the types of clauses in which they are found to be most prevalent.

5.2.1.1 Patterns of *BE* Overgeneration

Overgeneration found in L1-Malay ESL sub-corpus can be divided into two main categories; *BE + bare V* and *BE + Ved*. Occasionally *BE* is also found to be overgenerated before a main verb inflected with 3rd person *–s* (*BE + Vs*) and before modal verb (*BE + modal + V*). All the four patterns are discussed in this section.

BE + bare V overgeneration

Overgeneration of *BE* before a base form verb as recorded by previous studies (Arshad & Hawanum, 2010; Ionin & Wexler, 2001; Wee, 2009) is also found to be common in the L1-Malay learner data in this study. The construction is believed to be the outcome of tense and/or agreement marking (Ionin & Wexler, 2001; Ionin & Wexler, 2002; Lardiere, 1998), whereby *BE* is utilised as the mechanism to mark either one or both of the grammatical features. It is found that *BE + V* construction in the L1-Malay learner

sub-corpus to be utilised mainly as the marker for agreement. As can be seen in extracts (95) to (99) below, the subject verb concord consistently suggests that *BE* is used to mark agreement, consistent with the supposition postulated by Ionin and Wexler (2001, 2002) and Lardiere (1998).

(95). If it ***is happen***, it also may give a poor picture to the employer. B0039-05

(96). I also heard from the senior student who ***is already take*** the industrial training, they said that some of the company didn't give them opportunity to learn more about the work relate to their courses. B0037-05

(97). For the first week, they will attact people that customer will get more money if they ***are join*** that offer. C0007

(98). There are student who think that they ***are study*** just to pass the exam. B0039-05

(99). The medias ***are always show*** the luxurious and exclusive lifestyle and this can make someone who doesn't have money to acquire something luxurious to dream about it. I0009

As for *BE* inserted as a marker for tense feature (Arshad & Hawanum, 2010; Wee, 2009), the data in this study could not be used to verify the supposition, firstly due to the limited use of the past tense *BE* (*was, were)* in the data. Secondly, *was/were + V* construction is not utilised consistently. For instance in extract (100), the learner inserted *was* before only one of the three verbs (*was died*). If *was* is used by the learner as an indicator for past tense, then the same insertion pattern should also be applied to *want* and *came*. Thirdly, there is also no consistency in the forms of the lexical verbs proceeding *BE*. As can be observed in extracts (100) and (101) *was* is inserted not only before bare verbs (*was happen*), but also before inflected lexical verbs (*was died, was happened*) and more importantly they occur alongside one another e.g. *was happened* and *was happen* in extract (101). Due to these three reasons it is difficult to ascertain if *BE* is also used by the learners to mark the tense feature.

(100). The man **want** to get a money from the women, suddenly a few people **came** to helped that women. The man ***was died*** at that place because was bite by a few people helped that women. F0096

(101). Firstly, we can look the case corruption on this world. Let us see why it **is *happened***. Simply it ***was happen*** because the money. This ***was occurred*** because they need money to stay up in the foreign country. I0011

BE + Ved overgeneration

Past studies reveal a distinct pattern of *BE* overgeneration (*BE + Ved/en*) involving a subclass of intransitive verb; unaccusatives, such as *was sunk* and *was happened* (Balcom, 1997; Ju, 2000; Oshita, 2000; Yip 1995), which owing to its similarities to the English passives is termed overpassivisation (Ju, 2000, Oshita, 2000; Yip, 1995). Overpassivisation according to researchers is a product of either learners' confusion of the English passive (Oshita, 2000) or their inability to conceptualise the agent of unaccusative verbs (Ju, 2000; Yip, 1995). There are some instances of overpassivisation-like errors involving unaccusatives in the L1-Malay data and most often they involve the verb *happen* as exemplified in extracts (102) to (104) and occasionally several other unaccusative verbs such as *increase* (105) and *change* (106).

(102). Bank and internet robberies, snatch thefts, bribery, monetary speculations, terrorism, war after wars, increases in inflation rate, high rate of unemployment and so many such like, ***are happened*** because of money. B0198

(103). From this, we can conclude that, many *falling* (failing) marriage ***is happened*** because they have no time together. C0019

(104). Firstly, we can look the case corruption on this world. Let us see why it ***is happened***. Simply it ***was happen*** because the money. This ***was occurred*** because they need money to stay up in the foreign country. I0011

(105). Patients with bronchial problems and asthma ***are increase*** in number. S0034-05

(106). Every second technology ***were change***. S0023-05

Nevertheless, not all the overgenerated unaccusatives are inflected, some are left unmarked as can be seen in extracts (104), (105) and (106) with *was happen*, *are increase* and *were change* respectively, which could mean that learners were not attempting to passivise the verbs. Extracts (105) and (106) could also be impaired progressives, as both constructions transmit progressive meaning. The lexical semantic of the verb *increase* in the context of extract (105) for instance clearly refers to *growing* (with reference to the number of patients with respiratory problems). The overgeneration *was happen* that occurs alongside *is happened* (104) points to the confusion learners are experiencing not only with unaccusative verbs, but also with English IP system. As discussed in the previous section, the suppletive inflection *was* (e.g. *was happen* and *was occurred*) could be employed to check the tense feature. However, due to the inconsistency in the overgeneration construction, whereby *occur* is inflected but *happen* is left bare (extract 104), it is difficult to conclude that *BE* is used to check the tense feature in this study.

Another important finding with regard to overpassivisation-like overgeneration is that this construction involves not only unaccusative verbs, but also unergative verbs and transitive verbs as exemplified in extracts (107) and (108) respectively. The findings suggest that even though learners may experience difficulty realising the deep structure of unaccusative verbs, the overpassivisation-like overgeneration is not, however, the sole product of confusion with unaccusative verbs since they also involve unergative and transitive verbs. Transitive verbs are agentive in nature, they take subjects that assume the role of agent and hence they can be passivised. This finding suggests that overpassivisation-like overgeneration is not entirely driven by the lexico-semantic properties of the verbs, but interplay of other factors which could include faulty application of syntactic rules (NP movement in passive voice) and impaired understanding of English IP system.

(107). However, after the many of years, people ***are learnt*** to use money in their daily life. F0053

(108). Are the role of society in Malaysia ***are only censored*** the visual explicity in order to not tainted the society minds, I say reading can effect on par or rather more, to the society minds. A0001-05

BE + Vs overgeneration

Another interesting *BE* overgeneration pattern is *BE + Vs*, where *BE* is overgenerated before a lexical verb inflected with 3rd person singular *–s* morpheme as shown in extracts (109) to (112) below:

(109). It ***is depends*** on oneself to make him/her have the skills. (B0090)

(110). Learning process in university ***is begins*** with theoretical to introduce students to the roots of their field of study. (K0080)

(111). And the most important things that we as a student should and must change our mind set that the university degrees ***are only helps*** us to find the works but do not prepare students totally for the real world. F0084

(112). Other wise, in Malaysia, many people most of them among a rich people ***are says***, money is important. F0096

This construction suggests two possibilities; firstly, learners are double marking agreement by using both suppletive and affixal inflections and secondly the suppletive inflection is inserted to mark agreement feature, while the affixal inflection performs the task of checking the tense feature. The first possibility seems to only apply to extracts (109) and (110), whereby the singular suppletive inflection (*is*) is used concurrently with 3rd person singular morpheme *–s* in marking singular agreement. As for the second possibility, *BE* could be used as the agreement marker in all four extracts (109) to (112), as there is no variability in subject-verb concord in all the extracts, while affixal inflection could be utilised by the learners to check the tense feature as all the extracts are expressing present time.

BE + modal + V overgeneration

The L1-Malay learner data also reveal another unique overgeneration construction, whereby *BE* is inserted before a modal verb phrase as shown in extracts (113) to (116) below:

(113). So, when they are coming to interview, they ***are cannot speaking*** well with others. I0006

(114). Since these group of people really needs money, they ***are ought to kill*** others, robbing, rapping and *commiting* suicide. B0213

(115). Sometime, they ***are also can made*** a dangerous person like a murder, thief, pick *poket* and others. F0071

(116). I think, when they ***are joined*** this "*khidmat Negara*" they ***are can meet*** so many people from our country and they ***are can get*** many friends, finally they ***are can make*** a new relationship by each other. F0090

The subjects and *BE* concord in the extracts above seems to suggest that the suppletive inflection is employed to mark agreement consistent with its tendency in the overgeneration instances involving lexical verbs. It is observed that the modal verb phrases are used appropriately, except for extract (113) in which the present progressive *-ing* is used after modal auxiliary instead of its base form (*speak*).

The analysis has also unravelled another rare and unique overgeneration pattern involving modal auxiliary, whereby finite and non-finite *BE* are concurrently overgenerated. The finite *BE* is inserted before the modal verb phrase consistent with other patterns of overgeneration, while the non-finite *BE*, in this case infinitive *be* is inserted after modal auxiliary as exemplified in extracts (117) and (118) below:

(117). This situation ***is*** also ***will be happened*** in the real world when the student goes out to work. A0009-05

(118). So, they ***are must*** and ***should be collect*** many of money *untill* (until) the money make one of root of all evil. F0094

At an instance the constructions resemble modal in passive voice (*modal + be + Ved*) as in "*Each interpretation **can be seen** generally to flow through the abbreviated text as a whole*" (Biber et al., 2002, p. 184). Nevertheless, a closer look at the construction reveals that the infinitive *be* is inappropriately inserted, hence created the impaired modal passives exemplified in the extracts above. It is important to note that overgenerations involving modal verbs are very limited in the data.

5.2.1.2 Overgeneration of *BE* and Syntactic Environments

This section presents and discusses the possible influence of the syntactic environments on *BE* overgeneration instances. The constituents analysed include the type of subjects and the presence of intensifiers.

5.2.1.2.1 Overgeneration and Type of Subjects

Several studies on ESL learner variability in the use of *BE* especially involving omission of *BE* associated omission to the types of subject NP (Tode, 2003, 2007; Herat 2005; Wilson, 2003; Pine et al., 2008). The same pattern, however, is not evident in the overgeneration instances in the L1-Malay learner sub-corpus. As already attested by the quantitative findings, *BE* is found to be overgenerated at almost the same frequency after noun and pronoun subjects, suggesting that overgenerations are not influenced by the types of subjects. Extracts (119) to (121) are samples of overgenerations involving noun subjects, while extracts (122) and (123) are the samples involving pronoun subjects.

(119). Everyday we heard that <u>crime</u> ***is always happens***. B0034-05

(120). There are one way communication so that lecturer do not know <u>the students</u> ***are understand*** or not. B0038-05

(121). When <u>Gulf War</u> ***was begin***, the world economic also become down. B0016-05

(122). … doing the bad things to get money, <u>this</u> ***is means*** use money to get more money. B0007

(123). They ***are only take care*** about themselves, ignore other people because don't want to threat their safety. B0043-05

The next step is to examine the subjects in terms of their number (singular versus plural). Overgeneration involving noun subjects are observed to occur slightly more often with singular subjects than plural subjects, consistent with the quantitative findings that reveal slightly higher frequency of overgeneration of *is* with noun subjects. Extracts (124) and (125) exemplified the instances of overgeneration involving singular nouns (underlined), while extracts (126) and (127) are samples of overgeneration involving plural nouns.

(124). Job world ***is required*** not the typical one but person who can generate the profit for the firms/*goverments* in facing competitive world nowadays. E0040

(125). ...that the lyric is trying to convey the meaning of the evil ***is*** always ***happened*** because of money. FP0056

(126). Many of people ***are agree*** that the degree will make their life *comfortability* in their future. B0068

(127). University degrees ***are emphasis*** only on theoretical concepts has been debate a couple years ago. B0136

As for overgeneration involving pronoun subjects, plural pronouns are observed to instigate more instances of overgeneration than singular pronouns. Extracts (128) to (130) exemplified instances of overgeneration involving singular pronouns, while extracts (131) to (133) are samples involving plural pronouns. With regard to the types of pronouns, singular pronouns comprise mostly the 3rd person *it*, while plural pronouns are made of predominantly the 3rd person *they*.

(128). This ***is always happen*** among the wealth family especially who was living in a big city. B0054

(129). For student, they have a lot of knowledge that they should know, so that they only memorize everything, but when it ***is come*** to the real situation, they cannot do anything. B0041-05

(130). It ***is also gain*** a Prime Minister, Abdullah Badawi attention. B0136

(131). For student, they ***are study*** hard only to get the degree and not to get an education. B0041-05

(132). We ***are slowly learn*** and ***experience*** as much opportunity as we can. B0070

(133). It is different than polytechnic student. They ***are already has*** a working skill and theoretical too. E0015

Nevertheless, it is difficult to ascertain if there is any association between the subjects and overgeneration. In general, findings from the qualitative analysis suggest no association between the types of subjects to overgeneration instances.

5.2.1.2.2 Overgeneration and the Presence of Intensifiers

Overgeneration is also analysed in relation to the presence of intensifiers in particular degree adverbs. The presence of degree adverbs such as *also, always, only* and *still* as exemplified in extracts (134) to (137) below is common in the data, however, they are not very persistent to be considered as an important determinant to influence overgeneration instances.

(134). It ***is also gain*** a Prime Minister, Abdullah Badawi attention. B0136

(135). This ***is always happen*** among the wealth family especially who was living in a big city. B0054

(136). They ***are only take care*** about themselves, ignore other people because don't want to threat their safety. B0043-05

(137). Even what do we want to eat today ***is come*** from our imagination…So, in conclusion, we can say that the place for our imagination ***is still exist*** in this science, technology and industrialization area. F0106

5.2.1.3 Overgeneration of *BE* and Syntactic Complexity

Overgeneration instances in the L1-Malay learner data occur in multiple sentence patterns; from simple to complex sentences. This goes to show that it is not syntactically bound. Learners tend to use complex sentences in their writings and this may have contributed to more frequent instances of overgeneration involving complex

sentences. Nonetheless, the analysis has also unravelled instances of overgeneration in simple sentences suggesting that overgeneration is not constrained by a single type of clauses. The extracts repeated below are samples of overgeneration occurring in simple sentences (extracts 138-139), compound sentences (extracts 140-141) and complex sentences (extracts 142-144).

(138). Most of them ***are graduated*** in degrees. (E0046)

(139). It ***is depends*** on oneself to make him/her have the skills. (B0090)

(140). The statistic have proof it and the case ***are increased*** in recent year. E0014-05

(141). The evil ***are liked*** this people and the God's a very angry with this people. F0098

(142). And the most important things that we as a student should and must change our mind set that the university degrees ***are only helps*** us to find the works but do not prepare students totally for the real world. F0084

(143). Job world ***is required*** not the typical one but person who can generate the profit for the firms/goverments in facing competitive world nowadays. (E0040)

(144). I'm sure for those who ***are actually think*** that the practical learning is just one of the pre-requisite for graduates, they will not really take it the practical as a serious learning. B0082

5.2.2 Omission of *BE*

Past studies have recorded omissions of *BE* as one of the major types of errors among Malaysian ESL learners (Arshad & Hawanum, 2010; Maros et al., 2007; Wee, 2009). The same tendency is also observed in the essays of less proficient learners in this study. The following sub-sections present and discuss *BE* omissions in relation to the patterns they take, the influence of the syntactic environments and the syntactic complexity in which they occur.

In general, omissions of *BE* in the L1-Malay data can be divided to two major categories copula *BE* omission, which involves omission before a subject complement

(*subject* + ∅ + *complement*) and auxiliary *BE* omission, which refers to omission preceding a lexical verb (*subject* + ∅ + *Ved/Ving*). The symbol ∅ is used to indicate the missing *BE*. Each category is discussed further in the subsequent sections.

5.2.2.1 Omission of Copula *BE*

The analysis of copula *BE* omissions include all instances of *BE* missing before a subject complement as exemplified in extracts (145) to (147) below. Note that the subjects are bold and italicised, while the complements are underlined.

(145). Only if ***it*** ∅ only <u>on National television or radio channel</u> compare to the wide range of media. A0001-05

(146). ***It*** ∅ <u>same</u> as when we become a leader in a job. We can give a instruction easily to our workers. B0068

(147). ***That*** ∅ <u>why the censorship is very important in our society</u>. F0087

5.2.2.1.1. Copula BE Omission and Syntactic Environments

Herat (2005) in analysing *BE* variation in the spoken data of Sri Lankan speakers of English analysed zero *BE* in relation to the types of complements (noun phrase, adjective phrase, present participle and past participle), types of subjects (personal pronoun, other pronouns) and preceding phonological environment (whether the phoneme before *BE* is a consonant or vowel). *BE* omissions in this study are also examined in relation to the grammatical environments identified and analysed by Herat (2005), which include the types of complements and types of subjects. The preceding phonological environment was excluded from the analysis as this study focuses mainly on the written register, thereby eliminating the need for analysing the preceding phonological environment. In addition, this study has also included the presence of intensifiers (degree adverbs and negation *not*) preceding null *BE*, which Lee and Huang (2004) identified to be persistent constituents in the omission cases by L1-Chinese learners they examined.

5.2.2.1.1.1 Copula *BE* Omission and Type of Subjects

The findings of the quantitative analysis reveal higher occurrences of omissions before noun subjects compared to before pronoun subjects. Extracts (148) to (150) below are samples of omissions following noun subjects. There is also a strong tendency for *BE* to be absent after plural nouns as exemplified in extracts (149) and (150). Tode (2003, 2007) explained that plural noun subjects would generally pose more problems since learners would be required to ascertain the number of the subjects before they can determine the morphological forms of *BE*.

(148). *Co-riricular* (Co-curricular) activities ∅ also important to the student for get a job. B0039-05

(149). It is because fresh graduates ∅ not good enough to work with the company? B0070-05

(150). Maybe, this is because of the system in university nowadays, lecturer ∅ only a student. B0041-05

In contrast, pronoun subjects are believed to pose lesser difficulties to learners as plural formation of pronouns would not require any suffixation. Singular pronoun *it* for example is replaced by plural *they*, making it easier for the learners to process them as formulaic sequences, as specific pronoun will always be sequenced with specific morphological form (e.g. *it is..., they are..., you are...*) (Wilson, 2003). Nonetheless, omissions involving pronoun subjects as exemplified by extracts (151) to (153) are still found quite consistently in the data, suggesting that although they occur less frequently than after noun subjects, they still pose some difficulties to at least some of the learners.

(151). …it ∅ not enough if we want to more improve. E0029

(152). They ∅ afraid to steal time to enjoy with friends and doing other things. E0031-05

(153). Although we ∅ always busy with our study or our working, when we go home, we must dreaming to rest our body. E0073

5.2.2.1.1.2 Copula *BE* Omission and Subject Predicates

Platt and Weber (1980), reported that in the spoken Malaysian English, *BE* was often absent before adjectival, nominal and locative predicates. Herat (2005) in his investigation concluded that the Sri Lankan English speakers exhibited greater tendency to omit *BE* after adjectivals than other types of subject predicates. L1-Russian learners (Unlu & Hatipoglu, 2012) were also observed to show similar *BE* omission trend. Even though omissions were not considered as a major problem to the L1-Russian learners, but when they did occur, they were often followed by adjectivals. Lee and Huang (2004) also reported higher percentage of omissions before adjectivals in the data of L1-Chinese learners they analysed. These findings suggest that realisation of *BE* to an extent could be determined by the types of complements. In general, copula *BE* omissions in the L1-Malay learner data occur mainly before adjective predicates (*BE + AP*) as shown in extracts (154) to (157) below:

(154). Such as take them to work in their company but give a salary that ∅ *not suitable* with their hard work. B0005-05

(155). Public speaking skill ∅ *also important* in the real world. B0068

(156). It can carry more people compare to bus or taxi and also it ∅ *so fast*. C0008

(157). When rich parents ∅ *too busy* making money they didn't realise that they fail to give their children enough love and care. B0117

Other than adjective predicates, the absence of *BE* in the L1-Malay learner data also occurs considerably consistent before nominal predicates (*BE + NP*) as exemplified in extracts (158) to (161) below:

(158). Well, the money ∅ *not source* to all evil but source to *peacefull*. E0075

(159). Media mass ∅ *one of the factor influens* teenagers to get involves in wrong activities, media mass *expecially* television, crime *influens tenegers* do wrong activities. H0016

(160). Then there ∅ *also many program* that the government offer to the graduates like KPLI and so on. B0081

(161). In sum, money ∅ *just a self-valued paper* designed by humankind; thus we the humankind, ∅ *the ones* who ∅ responsible to use the money appropriately and rightfully. B0198

Contrary to the findings of Platt and Weber (1980), who reported the absence of *BE* before locative predicates as in *And my brother Ø also in Kedah* (Platt & Weber, 1980, p. 74), the omission of *BE* preceding locatives is almost non-existent in the L1-Malay learner data. In general, overt copula *BE* in the data is rarely complemented by locatives, consistent with the corpus findings of Biber et al. (1999), who reported rare instances of copula *BE* complemented by locatives in academic prose.

5.2.2.1.1.3 Copula *BE* Omission and the Presence of Intensifiers

The qualitative analysis reveals that it is common for null *BE* complemented by adjective predicates to be modified by either negation *not* as in extract (162) or degree adverbs such as (*also, so, too*) as in extracts (163) to (165) below:

(162). Such as take them to work in their company but give a salary that ∅ *not suitable* with their hard work. B0005-05

(163). Public speaking skill ∅ *also important* in the real world. B0068

(164). It can carry more people compare to bus or taxi and also it ∅ *so fast*. C0008

(165). When rich parents ∅ *too busy* making money they didn't realise that they fail to give their children enough love and care. B0117

Similar to the condition with omissions before adjective predicates, the realisation of *BE* before nominal predicates seems to be affected by the presence of negation *not* (166) and degree adverbs (extracts 167-168).

(166). Well, the money ∅ *not source* to all evil but source to peacefull. E0075

(167). Then there ∅ *also many program* that the government offer to the graduates like KPLI and so on. B0081

(168). In sum, money ∅ *just a self-valued paper* designed by humankind; thus we the humankind, ∅ the ones who ∅ responsible to use the money appropriately and rightfully. B0198

In general, the presence of degree adverbs seems to exert some influence in the realisation of *BE* in the L1-Malay learner data. Extract (169) below provides a good example of such influence, whereby *BE* is overt in the absent of a modifying adverb, but is covert in the presence of a degree adverb (*only*).

(169). In modern society, we can see very clearly how money makes people especially teenagers being spoilt, old people ***are*** abundant, parents become greedy… They ∅ ***only*** concern on their needs. FP0039

Nonetheless, compared to negation *not*, the instances of null *BE* complemented by adjective and nominal predicates modified by degree adverbs appear to be more common in the data. This is perhaps due to the limited use of negations in the data.

5.2.2.1.2 Copula BE Omission and Syntactic Complexity

The absence of copula *BE* is observed to also occur mostly in complex sentences (extracts 173-174), quite a number of instances are found in compound sentences (extracts 171-172) and occasionally in simple sentences (extract 170). In general, omission can occur in all three major sentence structures, however, due to the learners' preference to complex sentences makes the instances of omissions in complex sentences appear to be more prominent. The finding suggests that syntactic complexity is not a constraint to *BE* omission.

(170). They ∅ only concern on their needs. FP0039

(171). It can carry more people compare to bus or taxi and also it ∅ *so fast*. C0008

(172). Well, the money ∅ not source to all evil but source to *peacefull*. E0075

(173). If this group of people do not perform and apply *rasional* thinking, they might be get into doing something that ∅ out of mind. A0005-05

(174). When rich parents ∅ too busy making money they didn't notice that they fail to give their children enough love and care. B0117

5.2.2.2 Omission of Auxiliary *BE*

Past studies (Herat, 2005; Maros et al., 2007; Muneera & Wong, 2011; Platt & Weber, 1980; Unlu & Hatipoglu, 2012) have all attested the tendency learners have to omit *BE* more frequently before present participle *–ing* (*BE + Ving*) and past participle *–ed* (*BE + Ved*), thus, producing impaired progressives and passives respectively. The subsequent sections present and discuss the findings of *BE + Ving* and *BE + Ved* omission patterns found in the L1-Malay data.

BE + Ving omission

The analysis of *BE + Ving* omissions involves firstly determining whether the constructions express progressive aspect. Progressive aspect is used to express actions or activities in progress at a particular time either in the past or present time and usually the actions progress for a limited period of time. It can also have future time reference to describe actions or events that are going to take place in the future (Biber et al., 1999).

The analysis of *BE* omissions proceeding *Ving* reveals that they mainly express actions or events that are in progress in the present time. This is usually indicated with the use of adverbial of time such as *nowadays* and *always* as exemplified in extracts (175) to (177). Even when the adverbials of time are not present, the context of the sentences provides clue/s that the *Ving* verb is utilised to refer to an action that is in progress in the present time, for example in extract (178) the verb *using* is used to refer specifically to the system the university is currently employing. The use of communication verbs such as *talk* (179), which according to Biber et al. (1999) is a type of verbs that occurs 50% of the time in progressive aspect, also provides reference to the present time.

(175). ***Nowadays***, we ***∅ using*** a lot of money to get make sure the best living we have. E0014-05

(176). This is because he ∅ ***always wasting*** his money on buying somethings. E0030

(177). In this era of high-tech age of technology and industrialization, people ∅ ***always rushing*** everywhere to get anything such as money and leisure and perhaps no time to be daydreamer. B0112

(178). Furthermore, it is also depends on the system that the university ∅ ***using***. E0014

(179). So when we ∅ ***talking*** about that, we are show about value of student who are grade, student now not have a good *knowlegd* about real world. I0001

Other than expressing action in progress in the present time, there are also occasional instances of *Ving* occurring after null *BE* that are used to express actions that were in progress in the past as can be seen in extract (180). In general, the use of past tense is very limited in the data, making impaired past progressive as exemplified in extract (180) below very rare.

(180). While they were struggling to seize one of the pedestrian's handbag, the pedestrian (woman) ∅ ***yelling*** for help. C0024

Progressive aspect can also be used to describe future actions. Nevertheless, analysis of impaired progressives in the data reveals no such usages. In general, the learners are found to use of *Ving* mainly to express actions in progress in the present time.

The data also reveal some instances where the *Ving* are inappropriately used to express actions that are non-progressive. These instances could be the outcome of learners' confusion of the formation the negatives of action verbs, which can be seen in extract (181). The verb phrases *not teaching* and *haven't* in extract (181) are best replaced with the phrase *do not teach* and *do not have* respectively. In addition, contextually the verb *teaching* does not refer to an action that is in progress, besides there is also no signposting to signal any progression taking place.

(181). Some lecture ∅ ***not teaching*** very well, because they ***haven't*** more knowledge and expirience. H0009

BE + Ved omission

Omissions of *BE* in passives are considerably common among L2 learners, consistent with the findings of previous studies (Herat, 2005; Maros et al., 2007; Muneera & Wong, 2011; Platt & Weber, 1980; Unlu & Hatipoglu, 2012). Omissions of auxiliary *BE* passive voice are mainly characterised by *BE* omitted before verbs inflected with *-ed* or *–en.* Passive voice is characterised by (i) auxiliary *BE + Ved* formation (Biber et al., 1999), (ii) subject-object inversion (the active subject becomes the passive agent and the active object becomes the passive subject), and (iii) the introduction of preposition *by* before the agent (Biber et al., 1999, Quirk et al., 1985). The *by*-phrase is, however, an optional element. It is only required when the agent is specified and in the case when the agent is not specified, the *by*-phrase is not required (Biber et al., 1999).

It is found that *BE* omissions before lexical verbs inflected with *–ed* participle in the data do possess the characteristics of English passive especially in terms of subject-object inversion and they sometimes can be easily identified by the presence of *by*-phrase. These impaired passives comprise both short passives as in extracts (182) to (184) and long passives as in extract (185). Nonetheless, auxiliary *BE* omissions in the data tend to occur comparatively more frequent in short passives than in long passives. As mentioned earlier short passives are generally more common in academic prose (Biber et al., 1999), which in this case is also reflected in the omission instances in the L1-Malay learner data.

(182). Dreaming and imagining ∅ ***usually refered to*** the artist or people who ∅ ***involved*** in the art industry either performance, or illustration. K0001-05

(183). Most universities ∅ ***already filled*** with the lectures which are professor-centered. K0041

(184). In my opinion, censorship role in our country does not affect much in as good ways, these is because not only Malaysia ∅ ***known*** as the toughest censorship system but it ∅ ***also known*** as the biggest pirated VCD producer. A0008

(185). If we look to another development country likes Indonesia and Thailand, they ∅ ***also dominated*** by science, technology and industrialization. E0029

BE + V omission

Omission involving passive constructions is also found to take *BE + V* structure, whereby the main verb is left bare as exemplified in extracts (186) and (187). These extracts are clearly in the passive voice as characterised by subject-object inversion and the presence of *by*-phrase. In this case, learners appear to be unsure of the formation of passive voice that would require the use auxiliary *BE* with the main verb inflected with *–ed* participle (*BE + Ved*). Nevertheless, null *BE* preceding bare verbs in the passive voice are very few in the data, suggesting that they are not a prominent feature of the L1-Malay learner language in this study. It is also noted that extract (187) also involves omission of non-finite infinitive *be* after the modal verb *will* (*will be included*).

(186). Students must involve in projects or programmes that ∅ ***conduct*** by the university to increase their value. B0038-05

(187). Usually, any activities that ∅ ***participate*** by student will ∅ ***included*** in resume. B0039-05

5.2.2.2.1 Auxiliary BE Omission and Syntactic Environments

5.2.2.2.1.1 Auxiliary *BE* Omission and Type of Subjects

It has already been established through the findings from the quantitative analysis that both progressive and passive auxiliary *BE* omissions occur more frequently after noun subjects. Thus, the qualitative analysis provides the next level of analysis, whereby the subjects are examined in terms of their number (singular or plural).

Null *BE* is found to be preceded most often by plural noun subjects (extracts 188-191) consistent with the quantitative finding that reports higher frequencies of the omissions of auxiliary *are* compared to *is*. This means that there are lower occurrences of omission preceded by singular subjects (nouns and pronouns). As explained earlier plural noun subjects would generally pose more problems to learners since they would be required to ascertain the number of the subjects before determining the morphological forms of the proceeding *BE* (Tode, 2003, 2007). Omissions involving singular nouns similar to extract (192) are also available in the data, but as mentioned earlier they are less frequent than omissions involving plural noun subjects.

(188). There are <u>many seminar and workshops</u> that ∅ ***conducted*** by university for their student. B0038-05

(189). They play with <u>things</u> that ∅ ***not meant*** to be a toy or a toy is played in a different way. K0090

(190). <u>Dreaming and imagining</u> ∅ ***usually refered to*** the artist or people who ∅ ***involved*** in the art industry either performance, or illustration. K0001-05

(191). <u>Most universities</u> ∅ ***already filled*** with the lectures which are professor-centered. K0041

(192). In my opinion, censorship role in our country does not affect much in as good ways, these is because not only <u>Malaysia</u> ∅ ***known*** as the toughest censorship system but <u>it</u> ∅ ***also known*** as the biggest pirated VCD producer. A0008

Similar pattern is also observed with pronoun subjects, which also recorded more omissions after plural pronouns than after singular pronouns. It is also found that the 3rd person plural *they* is the most common type of pronoun subjects to precede null *BE* as exemplified in extracts (193) and (194). The occurrences of singular pronouns as subjects of null auxiliary *BE* are comparatively very limited as mentioned earlier. Extract (195) is a sample of null *BE* preceded by a singular pronoun *he*.

(193). Sometimes, for those who ∅ ***really cunning***, they ∅ ***just using*** the computer and pressing the certain number or button to transform money from others bank account into theirs. FP0042

(194). If we look to another development country likes Indonesia and Thailand, they ∅ ***also dominated*** by science, technology and industrialization. E0029

(195). This is because he ∅ ***always wasting*** his money on buying *somethings*. E0030

5.2.2.2.1.2 Auxiliary *BE* Omission and the Presence of Intensifiers

Omissions in the position before *Ving* are found to occur with or without adverbial modification. As shown in extracts (193) to (195) above, the verbs preceding null *BE* are modified by degree adverbs (*always*, *really*, *just*). The pattern is almost similar to copula *BE* omission reported by Lee & Huang, (2004). Nevertheless, it is also common to find omissions which involve no adverbial modification as exemplified in extracts (196) and (197) below. In addition, *BE +Ving* omissions are very rarely preceded by negation *not* as they would with degree adverbs.

(196). Since the world ∅ ***moving*** to globalization era, the crime also increase beyond the country like terrorism, smuggling and illegal immigrants. B0043-05

(197). So, that we can define a university as a place for students to further their studies in specific courses, before their ∅ ***working***. B0017-05

(198). This is because he ∅ ***always wasting*** his money on buying somethings. E0030

(199). In this era of high-tech age of technology and industrialization, people ∅ ***always rushing*** everywhere to get anything such as money and leisure and perhaps no time to be daydreamer. B0112

(200). Sometimes, for those who ∅ ***really cunning***, they ∅ ***just using*** the computer and pressing the certain number or button to transform money from others bank account into theirs. FP0042

Similar to the omissions before *Ving*, which are sometimes modified by degree adverbs, omission of auxiliary *BE* before *Ved* is also found to be modified by adverbs especially focusing adverbs such as *also* and *only* as shown in extracts (201) and (202) below. They are also found to be occasionally negated by *not* as can be seen in extracts (203) and (204). Adverbial modification seems to be quite persistent in the omission instances suggesting that the supply of *BE* could be constrained by the presence of adverbials as postulated by Lee and Huang (2004).

(201). If we look to another development country likes Indonesia and Thailand, they ∅ ***also dominated*** by science, technology and industrialization. E0029

(202). In my opinion, censorship role in our country does not affect much in as good ways, these is because not only Malaysia ∅ ***known*** as the toughest censorship system but it ∅ ***also known*** as the biggest pirated VCD producer. A0008

(203). Some cases ∅ ***not only done*** by local robber but also involved the immigrant. B0043-05

(204). When his heirs ∅ ***not satisfied*** about what they got, the conflic can exist where they will fight among them to get more than they got. E0024-05

5.2.2.2.1.3 Auxiliary *BE* Omission and Syntactic Complexity

The absence of auxiliary *BE* in the progressive aspect is observed to occur mostly in complex sentences (extracts 207-208) and only some can be found in simple (extract 205) and compound (extract 206) sentences. This condition is probably the result of the learners' preference for syntactically complex constructions. In addition, *BE* can be absent in the main clause (extract 207) and in the subordinate clause (extract 208), suggesting that omissions can occur in any type of clauses.

(205). Why ∅ people ***killing*** each other? C0019

(206). People ∅ ***just dreaming*** about a car, aeroplane and motorcycle, but today, all of it disexist. F0114

(207). Other people ∅ ***living*** in fear because they ∅ scared of robbery break in their home or kidnap for money. K0082

(208). From psychology view, the emotional of people who ∅ ***working*** longer is not so stable. C0019

5.2.3 Summary of Ungrammatical Use of *BE*

The qualitative analysis reveals that overgeneration of *BE* tends to occur in two distinct patterns; *BE + bare V* and *BE + Ved.* The former is believed to be the result of agreement marking, whereby the suppletive inflection (*BE*), as postulated by Ionin and Wexler (2001, 2002) and Lardiere (1998), is employed as the mechanism to mark agreement. Whereas, the latter is the results of interplay of three major factors, namely lexico-semantic confusion of unaccusative verbs, faulty application of NP movement in passive voice and impaired understanding of English IP system. As for the syntactic environments surrounding the overgeneration instances, the patterns of overgeneration appear to be unaffected by the types of subjects, the class of post-*BE* verbs and the presence of intensifiers. In terms of syntactic complexity overgeneration instances are not syntactically bound, hence they can occur in a wide range of clauses.

Omission of *BE* can be divided into two major categories, namely omission of *BE*-copula and *BE*-auxiliary. Covert copula *BE* is commonly complemented by either adjectivals or nominals taking *BE + AP* and *BE + NP* patterns respectively. They also have the tendency to be preceded by plural noun subjects. It is also observed that syntactic complexity is not a constraint to *BE* omissions, therefore, they can occur in all types of clauses and sentences.

Covert auxiliary *BE* takes *BE + Ving* and *BE + Ved* patterns, which refer to omissions in progressive aspect and passive voice respectively. Omissions of auxiliary *BE* in progressives mainly express progressive actions in the present time and commonly signalled by adverbial of time such as *nowadays* and *always.* As for missing auxiliary

BE in passive voice is mostly found in short passives. Both progressive and passive auxiliary *BE* omissions tend to occur more frequently after plural noun subjects and they are not constrained by the types of clauses.

The findings of the textual analyses of the major types of ungrammatical use of *BE* reveal that these ill-formed constructions generally do not impede communication. The non-target like *BE* constructions are found to be at times distracting, but not enough to severely distort meaning. Only when the ill-formed *BE* is combined with other linguistic impairments, such as wrong lexical choice, misspelling or wrong word order, communication is affected. Extract (209) below exemplies this condition. The prose was extracted from an essay of a Band 2 learner. It contains one well-formed *BE*-copula construction and three possible *BE* omissions (indicated by ∅). The overt and covert *BE* are observed to be employed in relatively simple constructions involving mainly *BE-adjective* (*money is useless*, *money ∅ very important*, *our life will ∅ bright and sunshine*). The prose also contains other grammatical errors, which include incorrect use of articles (*the money*) and prepositions (*towards our life*), wrong word choice (*sunshine* instead of *sunny*) and misspelling (*to life* instead of *to live*). In general, the incorrect use of *BE* does not distort meaning as the message on the importance of money is clearly conveyed through the first three sentences of the prose. Nevertheless, the limited range of vocabulary, repetitive and simple structures with noticeable grammatical mistakes have rendered the prose unclear or difficult to comprehend. Therefore, there is no doubt that these ill-formed constructions do affect the overall writing performance as they could directly or indirectly reduce the quality of learner essays.

(209). Today, life without money ***is*** useless. it ***∅ [is]*** because, money ***∅ [is]*** very important **towards** our life. Our life will ***∅ [be]*** bright and **sunshine** with money. Every day, people will go out to find **the** money. **The** People also will go out in the **mornig** and ***∅ [be]*** back **in the dark** to **collect** **the** money to get

luxurious life. **In the short word**, **what** we do need **the** money. So, most **of** people **very** need the money to **life**. F0095 B2

It is not surprising to find that prose containing grammatically incorrect *BE* constructions would also contain a host of other grammatical problems as demonstrated by extract (209) above, suggesting that learners' inability to construct well-formed *BE* constructions could root from deeper underlying acquisition problems.

CHAPTER 6

DISCUSSION

6.0 Introduction

This chapter discusses the major findings and provides explanation for the phenomena discovered concerning the grammatical and ungrammatical uses of *BE* by the L1-Malay learners and the influence of the syntactic environments on these uses. The discussion will be organised according to the research questions of the study.

6.1 The Overall Distribution of *BE*

This section discusses the major findings of the overall distribution of *BE* according to forms and functions. Some parts of the discussion will also include the findings obtained from the comparative analysis between L1-Malay learner sub-corpus and the NS learner sub-corpora. This section intends to provide answers to the first and second research questions of the study:

> RQ1: What are the similarities and differences in the use of *BE* in the essays compiled in MACLE and LOCNESS?
>
> RQ2: What are the distributional patterns for each form and function of *BE* in the essays written by L1-Malay learners in the Malaysian Corpus of Learner English?

6.1.1 Distribution of *BE* According to Forms

The analysis of all the forms of *BE* in the L1-Malay learner sub-corpus has revealed four major distributional patterns:

1. High Frequency of *BE*

There is an overall high frequency of *BE* forms in the L1-Malay learner sub-corpus suggesting that the use of *BE* is very common in the writings of the L1-Malay learners in this study. The same pattern is also observed in the data of American learners. The L1-Malay and American learners recorded almost similar ratios of occurrences of the finite and non-finite *BE* forms. The same pattern, however, is not found in the British learner data. They tend to use *BE* less frequently than the American and L1-Malay learners. Past studies have attested variations between American and British English in several grammatical aspects (Leech, 1999; Olofsson, 2004; Tottie & Hoffmann, 2006), and the current study provides further evidence to these variations with data from *BE* use.

As for the specific *BE* forms, *is* and *are,* and infinitive *be* are most frequently used by the L1-Malay learners. The same pattern is also observed in the NS learner sub-corpora. In terms of frequency of use, *is* is the most frequently used form in the L1-Malay learner sub-corpus followed by *are* and infinitive *be*, which is similar to the pattern found in the American learner sub-corpus, while in the British learner sub-corpus *is* records the highest frequency of use followed by infinitive *be* and *are*.

The past forms *was* and *were* record lesser occurrences across all the learner sub-corpora. The higher distribution of present *BE* forms could be related to the writing genre. According to Hunston (2002) past tense are mostly associated with narratives, while the present tense is associated with formal academic prose that deals with generalisations, observations, or descriptions. The essays compiled for MACLE and

LOCNESS are mainly argumentative, which would require the learners to argue against or for an issue and most of the time the arguments are conducted using the present tense.

The essay prompts might also determine the choice of verb tense employed by the learners. All the prompts in MACLE are phrased in the present tense (e.g. "*Crime does not pay"*, "*Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value"* and *"The role of censorship in society"*) and they concern general issues that are not time bound. Based on the prompts it would be natural for the learners to express their point of views using the here-and-now arguments with perhaps some references to past events, which has resulted in the ideas being expressed mostly in the present tense.

One important observation made in the overall use of *BE* and the specific *BE* forms is that the L1-Malay learners' patterns of the use of *BE* are very similar to the American learners, but not to the British learners. This could have been caused by more exposure to American contents i.e. textbooks or references written in the USA (printed and online), internet search engines operated with American English or television/internet programmes produced in the USA. Nevertheless, based on the findings of this study it could not be ascertained if these are the causes of the similarities, hence further research needs to be conducted to examine these aspects further.

2. High Frequency of Finite *BE*

Finite *BE* forms are used more frequently by the L1-Malay learners than non-finite forms. This is consistent with the overall pattern of use of *BE* by the NS learners, who also tend to use more finite *BE* forms. However, it is important to emphasise that the percentage of use of the finite *BE* forms by the L1-Malay learners is considerably higher than their NS peers.

Finite *BE* forms are expected to occur more frequently than the non-finite forms. Finite *BE* is multi-functional, it can be used as either a copular or an auxiliary and is realised not only in declaratives, negatives, interrogatives and in existential *there* and *it*-cleft constructions. Finite *BE* also has more inflections than non-finite *BE*. Including the contracted forms, finite *BE* has a total of eight inflections (*am, is, are, was, were, 'm, 's, 're*) compared to only three non-finite forms (*be, been, being*). These reasons could have contributed to the higher occurrences of finite *BE* in not only the L1-Malay learner sub-corpus, but also in the NS learner sub-corpora.

According to Ellis (1985), L2 learners especially those who are less proficient, tend to avoid non-finite *BE* constructions. Non-finite *BE* is often used in comparatively more complex structures, for instance after auxiliaries such as after *have* as in "*The import of live chickens* ***have been suspended*** *since July",* after auxiliary *BE* as in "*The prisoners* ***are being transferred*** *to a more secured correctional facility"* and after modals as in "*The building* ***will be demolished*** *very soon"* (Ellis, 1985, p. 237). Nevertheless, the ratios of non-finite *BE* in the L1-Malay learner sub-corpus are similar to that in the American learner sub-corpus, which implies that the L1-Malay learners' pattern in the use of non-finite *BE* can neither be linked to avoidance strategy, nor can it be directly associated with the learners' proficiency in English. Even though non-finite *BE* forms are used less frequently in the L1-Malay data, they are often used accurately in complex structures similar to those exemplified by Ellis (1985).

3. High Frequency of Grammatical Forms

One of the major findings of this study is that the L1-Malay learners produce significantly higher frequency of grammatical use of *BE* compared to the ungrammatical use; 90.3% and 9.7% respectively, suggesting very high competency in the use of *BE* among the learners. Evidence from the qualitative analysis shows that *BE* is used in structurally complex constructions to express complex arguments. It is also

used in the constructions of *it*-clefts, which are considered as more advanced copular construction (Hinkel, 2003). The qualitative analysis reveals that *it*-clefts are more common in essays written by learners with higher proficiency in the English language, who scored Band 5 and above in the MUET.

4. Low Frequency of Ungrammatical Use of Non-Finite *BE*

Another important finding is that non-finite forms are very rarely misused by L1-Malay learners. Even though they do not occur as often as the finite forms, they are most often used correctly. Schütze (2004) in his investigation of omission rate of finite and non-finite *BE* in the acquisition of *BE* among children acquiring L1 English found that non-finite forms were rarely dropped, unlike the finite forms. The researcher explained that omission of finite forms is the result of underspecified Tense resulting in the condition where verb requirement no longer exists and the presence of *BE* is, therefore, not required. In contrast, non-finite forms exist alongside a modal or auxiliary, Tense is realised on the modal or auxiliary, forcing the presence of *BE* when Tense imposes its V requirement. Non-finite forms do not carry tense or agreement features like finite forms, since these features are already realised by the auxiliaries or modals preceding them. As a result, non-finite forms do not have to undergo morphological transformation like the finite forms. This makes using non-finite forms easier as the confusion learners might have in realising the tense and agreement features is already eliminated. This could explain why non-finite forms are rarely used incorrectly compared to the finite forms.

6.1.2 Distribution of *BE* According to Functions

This section discusses the overall distribution of *BE* according to the functions that it performs. The discussion in this section concentrates mainly on the use of *BE* as a copular and an auxiliary.

1. High Frequency of Copula *BE*

In the L1-Malay learner sub-corpus, *BE*-copula constructions occur more often than auxiliary *BE*. The same trend also exists in both the British and American learner sub-corpora. Milton (2001) has also reported similar finding. He found that auxiliary *BE* is relatively rare in the interlanguage grammar of the Hong Kong ESL learners. He claimed that this was the result of learners' avoidance of the passive voice and aspectual verb forms (Milton, 2001). However, the same claim could not be made for the L1-Malay ESL learners in this study. The written register and writing genre could be factors contributing to the lower occurrences of auxiliary *BE*. Progressive aspect is mainly used to express ongoing events or a temporary state (Quirk et al., 1985); hence, it is more common in speech (Biber et al., 1999). The opportunity to use this aspect might not be available in argumentative essays as it would in other genres such as narratives. These could be the reasons why the occurrences of *BE* in progressives are relatively low not only in the L1-Malay learner sub-corpus, but also in the NS learner sub-corpora.

As for the proportion of use, the L1-Malay learners have recorded the highest percentage of copular constructions compared to the NS learners; 64% in the L1-Malay learner sub-corpus, 54% in the American learner sub-corpus and 46% in the British learner sub-corpus. Hinkel (2003) suggests that heavy reliance on *BE*-copula constructions among NNS is common and this renders the writings of the NNS learners simplistic and conversational. The same characteristics, however, are not found in the *BE*-copula constructions in the writing of highly proficient L1-Malay learners in this study. The qualitative findings show evidence of complex and oftentimes sophisticated use of *BE*-copula constructions. The prose often contains *BE* in complex constructions with the main clauses extended by one or more subordinate clauses. According to

Hinkel (2003), the employment of "accurate and extensive use of subordinate clauses contributes to a higher degree of text sophistication" (p. 276).

According to Hinkel (2003) NNS learners' heavy dependency on simple *BE*-adjective constructions has also resulted in writings to be inferior to NS writers. Even though predicative adjectives are common complements to copula *BE* in the L1-Malay data, they are used as parts of syntactically complex clauses. The adjective predicatives are frequently complemented by post-predicate complements, such as *that*-clause, *wh*-clause, *to*-infinitive clause, *ing*-clause or prepositional phrase (e.g. …*are* ***confident*** *that having a lot of money*...). The same tendency can also be observed with *BE*-noun constructions, which are also subjected to similar post-predicate complements (e.g. …*is* ***a multiracial country*** *consisting people of different races*…).

Corpus analysis findings have established that copula *BE* is a "much more frequent verb in academic prose than in conversation, newspaper and fiction" (Biber et al., 2002, p. 141). These findings suggest that the higher occurrences of *BE* in the L1-Malay learner data is a common feature in academic writing, thus, not a reflection of the learners' inadequacy in producing syntactically complex and sophisticated constructions as suggested by Hinkel (2003). The similarities in the distribution of copula *BE* between the L1-Malay learners and NS learners also suggest that the L1-Malay learners possess an almost native-like intuition in the use of *BE* in their writings.

2. Higher Frequency of Passives

Auxiliary *BE* in the L1-Malay learner data is used mainly in the constructions of passives (*BE + Ved*/*en*). The learners also tend to exhaust the non-finite *BE* forms in the construction of passives by using future passive (*modal +BE + Ved*) and perfect passive (*have + been +Ved*). This finding corroborates with the findings of past studies (Biber et al., 1999; Johns, 1997; Swales, 1990; Swales & Feak, 2012) that reported

common occurrences of passives in academic prose. Traditionally, in academic writing passive voice are considered more suitable than active voice and they are also regarded as requisite in the written genres of certain disciplines such as engineering and natural sciences (Swales & Feak, 2012). Therefore, it is not surprising for passives to be used more frequently in the L1-Malay learner data, which consists of entirely academic compositions.

The qualitative analysis of the L1-Malay learner data also reveals stronger preference among L1-Malay learners to use short passives compared to long passive. This finding is substantiated by Biber et al. (1999), who also reported extensive employment of short passives in academic prose. Short passives, according to Biber et al. (1999) are most often employed when the presence of human actor (agent) is not required or not important, thus, places the status to the topics or direct object of the corresponding active voice.

In addition, there is also a general inclination for the passives in the L1-Malay learner data to be realised in structurally more complex constructions, which often involve complex sentences extended by one or more subordinate clauses. Other than located in the main clauses of complex constructions (e.g. *In university, courses* ***are conducted*** *in such a way*…), passive clauses are also located in the dependent clauses of the complex constructions (e.g. ...*The expectation is* <u>*that the university leaders* ***are drawn*** *from the best brains in society*</u> *and they can play the integrative multiple roles of being*…).

According to Swales (1990), passive voice is highly conventionalised in academic prose. Johns (1997) added that passives are not only conventionalised grammatical features of formal research articles, but also in students' writings at the university levels. Hence, the consistent employment of passive voice by the L1-Malay learners demonstrates a greater awareness to academic writing conventions among the learners.

More importantly the accurate and appropriate use of passives, which according to Hinkel (2004) are grammatically, lexically and pragmatically complex, provides evidence that the L1-Malay learners generally possess advanced competency in the constructions of passives.

6.1.3 Summary of Overall Distribution of *BE*

Triangulating from the findings of the quantitative, qualitative and comparative analysis, the L1-Malay learners can be concluded to be highly proficient in the use of all *BE* forms. This is clearly reflected in the higher ratios of the grammatical constructions of *BE* in the learner data. The L1-Malay learners are also more reliant to finite *BE* constructions compared to the NS learners. Even though finite *BE* occurs more frequently across the learner sub-corpora, the percentage of employment of finite *BE* in the L1-Malay learner data is comparatively higher. Nonetheless, the overall distribution of the grammatical use of *BE* in the L1-Malay data with regard to forms and finiteness largely corresponds with the distribution of *BE* in the American learner sub-corpus. However, the quantity of use is not the yardstick used to determine the L1-Malay learners' ability in using *BE*, rather they are only used as a reference in highlighting the overall patterns of the use of the verb by the L1-Malay learners in this study.

In terms of functions, *BE* in the L1-Malay learner sub-corpus performs two major functions, namely as a copular or an auxiliary. Consistent with findings of Biber et al. (1999), copula *BE* is especially more prominent in the L1-Malay learner data compared to auxiliary *BE*. As for auxiliary *BE*, it is found to be used mainly in the constructions of passive voice. This tendency could stem from the need to conform to the academic writing convention, which considers the use of passives as requisite.

6.2 Patterns of Grammatical and Ungrammatical Uses of *BE*

This section discusses the patterns of the grammatical and ungrammatical uses of *BE* in the study. It aims to address the third research question of the study:

> RQ3: What are the patterns of the (a) grammatical and (b) ungrammatical uses of *BE* in the essays written by L1-Malay learners?

6.2.1 Patterns of Grammatical Use of *BE*

The discussion for the patterns of the grammatical use of *BE* is divided to two sub-sections. The first focuses on the patterns of grammatical use of finite *BE,* while the second concentrates on the patterns of grammatical use of non-finite *BE*.

6.2.1.1 Grammatical Constructions of Finite *BE*

This section discusses the grammatical use of finite *BE* forms in two main constructions namely, copula *BE* and auxiliary *BE*.

6.2.1.1.1 Copula BE Constructions

L1-Malay learners in this study are generally proficient in the construction of all *BE*-copula patterns (e.g. *BE*-nominal, *BE*-adjectival, *BE*-preposition). However, *BE*-nominal and *BE*-adjectival are used more frequently than *BE*-preposition. *BE*-copula constructions in the L1-Malay learner sub-corpus mostly make use of lexical noun (NP) or personal pronoun (PPN) as subjects and predicated mainly by adjective (AP) or noun (NP) phrases as summarised in (1) below:

1. (a) *NP + BE + AP*
   (b) *PPN + BE + AP*
   (c) *NP + BE + NP*
   (d) *PPN + BE + NP*

BE-copula constructions using definite or indefinite pronouns as the subjects or prepositional phrase, *that*-clause, *wh*-clause and infinitive-*to* clause as the subject

predicates are also found in the learner data, but they tend to occur less frequently than the constructions summarised in (1). According to Biber et al. (1999), it is common for academic prose to contain more *BE*-noun and *BE*-adjective structures as they have specific functions to perform in such texts. In *BE*-noun structure, the noun phrase is commonly used to characterise or identify a subject, while the adjective in *BE*-adjective structure is typically used with other complements to express intellectual claims (Biber et al., 1999).

It is also important to add that the constructions in (1) are simplified representations of the actual constructions found in the learner essays. The qualitative findings reveal that *BE* is often realised in structurally complex constructions. In the complex *BE* constructions, it is common for the predicates to be further expanded by another clause such as infinitive *to* clause or a phrase such as a prepositional phrase. It is also common for the *BE* to be utilised in the subordinate clauses, which are marked by a subordinator *that* or *wh*-word, by non-finite verb phrases introduced by *to*-infinitive, *-ing* participle or *–ed* participle or by subordinating conjunctions such as *although*, *before* or *because*. These subordinate clauses can either be nominal, adjectival or adverbial clauses. The findings from the qualitative analysis reveal that these complex structures are very common in the L1-Malay learner data. This is contrary to the findings reported in Hinkel (2002) that L2 learners tended to produce repetitive and overly simplistic *BE*-copula constructions. The use of all these complex structures in the L1-Malay learner data can be regarded as the markers of textual and structural complexities (Hinkel, 2002). These copular constructions also include negatives and interrogatives, which are also preceded by either a NP subject or PPN subject and complemented by either a NP or AP predicate as shown in (2):

2. (a) *NP/PPN + BE + not + AP/NP* - negative
   (b) *BE + PPN/NP + AP/NP?* - interrogative

Copula-*BE* negatives in the L1-Malay learner data are also more commonly realised in complex sentences. The qualitative findings show that the negative constructions tend to be complemented by nominal predicates that are often modified by post-modifiers such as a prepositional phrase or *to*-infinitive clause, which contributes to their structural complexity.

Interrogatives in the L1-Malay learner data consist mainly of yes/no questions using VS word order with *BE* as the operator as shown in (2b). According to Biber et al. (1999), interrogatives in academic prose most often have rhetorical purposes. Similarly, interrogatives in the L1-Malay learner sub-corpus are also used as rhetorical questions. They tend to perform three major functions, namely to direct the reader to the essay topic, enforce a main point and form persuasion.

The copular constructions in the L1-Malay learner data are also realised in existential *there* constructions. The subjects for these constructions are usually NP that can be categorised into indefinite or definite NP. The former includes the NP subjects that are general such as "*some students*" or "*little effects*", while the latter includes those that have specified numbers or volume such as "*three men*" or "*a kilogram of flour*". Existential *there* structures are normally expanded by adverbials (Biber et al., 1999), however, in the L1-Malay learner data that is not the case. Instead they tend to be expanded by post-modifying clauses, namely nominals, infinitive-*to* or relative clauses. The L1-Malay learners in this study tend to produce the following existential there constructions:

3. (a) *There +BE + Ind NP + Expansion*

   (b) *There +BE + Def NP + Expansion*

Finally, the L1-Malay learner data also include *it*-clefts, which according to Hinkel are more advanced and sophisticated *BE*-copula (Hinkel, 2003). *It*-clefts in the L1-Malay learner data occur in two major patterns as shown in (4):

4. (a) *It + BE +AP*

   (b) *It + BE +NP*

It is interesting to note that L1-Malay learners also produce relatively complex *it*-clefts, which often contain a wide range of attendant elements. Adjective and noun predicates are often modified by a prepositional phrase or infinitive-*to* clause, which are then subsequented by a relative clause (e.g. *In the fast changing world of technology,* ***it is important to realise*** [*that there is a very high possibility that they may end up in areas they are not trained for*]). *It*-clefts are used quite consistently in the L1-Malay learner data especially among the more proficient learners. Considering that *it*-clefts are advanced syntactic construction (Hinkel, 2003), the employment of the structure in the L1-Malay learner essays is a clear indication that the learners are capable of syntactically complex *BE*-copula constructions.

From the detailed examination of both the quantitative and qualitative findings, it is possible to conclude that the L1-Malay learners in this study are proficient in the formation of all the major *BE*-copula constructions. As mentioned earlier some copular constructions (negatives, interrogatives) occur less frequently than the others due to perhaps the writing genre or/and the written register (Biber et al., 1999).

6.2.1.1.2 Auxiliary BE Constructions

Auxiliary *BE* is used to either mark progressive aspect as in "*Chris* ***are living*** *there now.*" (Biber et al., 2002, p.163) or in the formation of passives as in "*We* ***are delighted*** *with the result.*" (Biber et al., 2002, p. 167). The following sub-sections discuss the

patterns of the grammatical constructions of auxiliary *BE* in progressive aspect and in the formation of passive voice in the L1-Malay learner sub-corpus.

6.2.1.2.2.1 Auxiliary *BE* in the Formation of Passives

The percentage in the use auxiliary *BE* in passives by the L1-Malay learners is almost similar to their NS learner peers. The L1-Malay learner sub-corpus records 12.4% occurrences of passives compared to 11.2% in the American learner and 8.5% in the British learner sub-corpora. Auxiliary *BE* is also used more frequently in the construction of passives than progressives. This finding is consistent with the findings of previous research on written corpora that reported more common occurrences of passives in academic prose than in any other genres (Biber et al., 1999; Swales, 1990; Quirk et al., 1985).

These passive constructions are most frequently preceded by NP subjects, followed by PPN subjects and all of them are formed with transitive verbs (Vt). They are most often agentless and may be complemented by a prepositional phrase as in (5a) or contain a *by*-phrase as in (5b) below:

5. (a) *NP/PPN + BE + Vt-en + PP*

(b) *NP/PPN +BE + Vt-en + by-phrase*

L1-Malay learners prefer short or agentless passives. Biber et al. (1999) report that short passives are six times as common as long passives and tend to be extensively employed in academic prose. They can be found in the main clauses of complex sentences expanded by a prepositional phrase or in the dependent clauses of complex sentences.

There are also long passives in the L1-Malay learner data, which tend to occur in complex sentences and dependent on the post-*BE* lexical verbs and the importance of the agent. Certain verbs such as *grip* (*We* ***are gripped*** *by fear* …) or *cause* (*...murders*

committed ***were mostly caused*** *by jealousy or by mentally-ill people*.) require the presence of an agent. Most often, the by-phrases are complemented by nominal dependent clauses, which places important focus on the agents. The long passives in the learner data are commonly constructed in this manner, suggesting that long passives are only used when learners need to draw the readers' attention to the agent.

The use of passives in the L1-Malay learner data suggests most importantly that the learners are adhering to and exhibiting greater awareness to the conventionalised use of passives in academic writing, where passive voice is expected in the projection of objectivity and detachment (Hinkel, 2004). Secondly, its varied and accurate use also signal that the learners are capable of syntactically complex *BE* constructions as passives due to its "complex grammatical, lexical and pragmatic features is very difficult for NNS learners to use correctly and in appropriate contexts" (Hinkel, 2004, p. 24).

6.2.1.2.2.2 Auxiliary *BE* in the Formation of Progressives

Progressive aspect is used in expressing activities or events that are in progress at some point in time, it can be used to describe events progressing in the present, past and future time (Quirk et al., 1985). According to Biber et al. (1999) and Hunston (2002), progressive is relatively rare in academic prose as its employment may impart conversational flavour to academic writing, which then renders it less effective. Swales (1990) in his seminal work on the structure of academic discourse stressed that progressive aspect is hardly ever available in academic writing. Hinkel (2004) in her study on the use of tense, aspect and passive in academic writing of NNS learners reported that progressive aspect was simply not employed in the learners' writings. Therefore, it is not surprising that the use of this aspect in the L1-Malay learner data is

relatively limited. There are only about 6% of overall occurrences of *BE + Ving* constructions in the L1-Malay learner data.

These progressive constructions are preceded most frequently by NP subjects and followed by PPN subjects. The post-*BE* verbs used more frequently in these construction are transitive (Vt) and unergative (Uer) verbs, with Vt used more often than Uer. The constructions of the progressives are summarised in (6) below:

6. (a) *NP/PPN + BE + Vt-ing + expansion*

   (b) *NP/PPN + BE + Uer-ing + expansion*

Interestingly, the qualitative findings reveal that progressive aspect is used by the L1-Malay learners as a means to build reader-writer interaction. Even though the use of progressive aspect may impart conversational flavour to compositions, its use is found to be an effective way to establish reader-writer interaction. Nevertheless, in general progressive aspect is not common in the L1-Malay learner data.

6.2.1.2 Grammatical Constructions of Non-Finite *BE*

This section discusses the grammatical constructions of non-finite *BE* forms, which include *be, been* and *being*. Non-finite *BE* constructions in the L1-Malay learner sub-corpus are relatively fewer in comparison to finite *BE* constructions. This is highly expected as non-finite *BE* includes only three forms (*be, been* & *being*). Subsequent sections discuss the patterns of use of each non-finite *BE* form.

6.2.1.2.1 Infinitive be

Infinitive *be* is used in the formation of simple future (*modal + be*), passives (*modal + be + Ved*) and progressive aspect (*modal + be + Ving*). Only the first two constructions are used most frequently by the L1-Malay learners. Approximately 40% is used in the construction of passives (*modal + be + Ved*) and almost 38% for the construction of the simple future tense (*modal + be*). These constructions are often preceded by either NP

or PPN subjects. *Modal + be* structure is often complemented by AP or NP predicates, while the passive constructions are followed by transitive verbs as shown in (7) below:

7. (a) *NP/PPN + modal + be + AP/NP*

(b) *NP/PPN + modal + be + Vt-ed + expansion*

The qualitative analysis reveals that infinitive *be* used in the construction of passives commonly performs two major functions, namely making suggestions or recommendations and expressing obligation or possibilities. Modal auxiliaries *should* and *could* are most commonly paired with infinitive *be* to produce these passive constructions such as *should be addressed/could be reduced*. *Should* and *could* are generally used in academic prose to mark logical necessity and logical possibility respectively (Biber et al., 1999).

Infinitive *be* employed in active voice in *modal + be* constructions is mainly used for making suggestions and expressing obligation. As hedging devices, modals like *could*, *would* and *should* are most often used by learners to down tone their argument and create a positive reader-writer relationship. Studies investigating the use of interactional metadiscourse in academic texts (e.g. Dana, 2008; Hyland, 2005; Hyland & Tse, 2004) have long noted that hedging has become conventionalised in academic writing and the use of these devices is encouraged as they help writers to adhere to the convention of academic writing (Dana, 2008).

6.2.1.2.2 Non-finite been

The form *been* is normally used in the formation of present or past perfect (*have/has/had + been*), perfect passive (*have/has/had + been + Ved*) and perfect progressive (*have/has/had + been + Ving*). The L1-Malay learners tend to have very restricted use of *been* and it is used mainly in the formation of perfect passive (*have/has/had + been + Ved*). Approximately 84% of *been* is employed in the formation of perfect passive constructions. Biber et al. (1999) explain that perfect aspect

and passive voice are both common in academic prose, hence, it is not surprising that learners are more inclined to use *been* in this combination. These constructions are preceded most frequently by NP or PPN subjects and the majority of them are followed by transitive verbs. The textual analysis of the perfect passive constructions in the L1-Malay learner data shows that they are used to refer to past events that have present relevance. Perfect passive construction occurring in the L1-Malay learner data is as shown in (8) below:

8. (a) *NP/PPN + have + been + Vt-ed + expansion*

6.2.1.2.3 Non-Finite being

The form *being* is mainly used in the formation of progressive passive (*BE + being + Ved*) to express actions that are in progress or incomplete in the present, past and near future. Consistent with the corpus finding of Biber et al. (1999), which reported rare occurrence of progressive passive in academic prose, progressive passive is also rare in the L1-Malay learner sub-corpus.

The limited use of progressive passive by the L1-Malay learners could be because they are not presented with the opportunity to use the structure in their essays as progressive aspect is generally rare in academic prose. Another reason could be due to avoidance strategy employed by the learners, as progressive passive is considered more complex than any other non-finite *BE* constructions. Nevertheless, it is interesting to note that learners who opt to use this structure tend to use it appropriately and it is mainly preceded by NP subjects and proceeded transitive verbs as shown in (9) below:

9. (a) *NP + BE + being + Vt-ed + expansion*

6.2.2 Patterns of the Ungrammatical Use of *BE*

This section recapitulates the major patterns of the two most distinct ungrammatical use of *BE* in the L1-Malay learner data, overgeneration and omission of *BE*.

6.2.2.1 Overgeneration of *BE*

Overgeneration of finite *BE* involves insertion of *BE* before a main verb resulting in an ungrammatical *BE + V* structure as in "*is take, are study"*. Overgeneration in this study is relatively persistent suggesting that it is among the problematic areas faced by some of the L1-Malay learners involved in the study. *BE* overgeneration has to be differentiated from overpassivisation which is another type of *BE* insertion (Hirakawa, 2006; Ju, 2000; Oshita, 2000; Park & Lakshmanan, 2007; Yip, 1994). Overpassivisation has a structure similar to a passive (*be + Ved/en*), and it mainly involves the use of unaccusative verbs such as *happen* and *fall*. In contrast, insertion of *BE* in the L1-Malay involves mainly transitive verbs (e.g. *take*, *study*) and very few involve unaccusative verbs and the post-*BE* verbs are often left uninflected. Due to these differences, insertion of *BE* in this study is termed overgeneration following the term used by Ionin and Wexler (2001) to refer to similar instances of *BE* insertion in their data.

The overgeneration instances are analysed in relation to the form and class of post-*BE* verbs, the type of subjects and the presence of intensifiers and auxiliaries. The followings are the major findings of *BE* overgeneration in the L1-Malay learner data:

1. Higher occurrences of overgeneration with transitive verbs

The quantitative analysis of the class of post-*BE* verbs reveals that overgenerated *BE* favours transitive verbs (49%) over unergative (28%) and unaccusative (15%) verbs (Vt>Uer>Uac). This finding is consistent with past studies (e.g. Fleta, 2003; Ionin & Wexler, 2001) that reported higher likelihood for *BE* to be inserted before active transitive and unergative verbs. This pattern is contrary to overpassivisation that occurs mainly with unaccusative verbs (Hirakawa, 2006; Ju, 2000; Oshita, 2000; Park & Lakshmanan, 2007; Yip, 1994). This finding suggests that lexico-semantic of

unaccusative verbs is not a major cause for confusion among the learners involved in this study.

2. Higher occurrences of overgeneration before uninflected active verbs

The analysis of forms of post-*BE* verbs in the overgeneration constructions reveals that about 62% are constructed with uninflected or bare verbs. ESL learners' choice of the infinitival form of the lexical verbs is not uncommon as this has already been attested for in previous studies (Arshad & Hawanum, 2010; Fleta, 2003; Ionin & Wexler, 2001; Lardiere, 1998; Lee & Huang, 2004; Unlu & Hatipoglu, 2012; Wee, 2009). According to Lardiere (1998) and Ionin and Wexler (2001), the finite *BE* in *BE + V* construction is treated as a default for marking agreement and/or tense, similarly as would non-finite forms in the place of finites in early child verb production (Prevost & White, 1999, 2000). The qualitative findings reveal consistent subject verb concord in the overgeneration constructions suggesting that the learners rely on their mastery of the suppletive inflection to mark agreement. The finding from this study provides empirical support to Lardiere (1998) and Ionin and Wexler (2001). Nevertheless, this study has only managed to provide support for the use of *BE* in instantiating agreement. The very limited occurrences of past tense *BE* in overgeneration instances do not permit the researcher to analyse the possibility of the verb to function as the marker for tense feature.

The consistent pattern of overgeneration instances in the L1-Malay learner data suggests that they could be intralingual errors, which reflect characteristics of rule learning such as faulty generalisation and incomplete application of rules of the target language (Richards, 1971). Similarly, Scovel (2001) identifies intralingual errors as those that have resulted from the confusion learners experience with patterns of the target language, regardless of how the target language patterns are different than the learners' L1 (Scovel, 2001). The utilisation of suppletive inflections as the marker for agreement

suggests that learners are overgeneralising the rules applied to auxiliary *BE* in progressive aspect and in passive voice, whereby the tense and agreement features are marked by auxiliary *BE* as in "*He **is listening** to music/ They **were listening** to music*". Arshad and Hawanum (2010) provided similar explanation concerning overgeneration found in the Malaysian learner data they analysed.

Unlu and Hatipoglu (2012) also attributed overgeneration instances in the Russian learner data they analysed to intralingual errors. They categorised the errors as misformation errors following the surface strategy taxonomy of errors proposed by Dulay, Burt and Krashen (1982). According to Unlu and Hatipoglu (2012), misformation errors suggest incomplete understanding and application of the rule of the target language, which in this case resulted in misused copula *BE*. Both error categories, overgeneralisation and misformation, attribute learner errors to incomplete or impaired application of the target language rules, therefore, not the outcome of direct interlingual transfer. In Russian *BE* is omitted as a rule in the present tense and Malay does not have copula-like verb, however, overgeneration errors documented by learners of both languages could not have been caused by interlingual transfer as overgeneration construction would not have been allowed in both Russian and Malay grammars.

Despite the higher percentage of the uninflected verbs occurring in *BE* overgeneration, some 38% of the overgeneration constructions occur with inflected lexical verbs. As mentioned earlier, previous studies (Hirakawa, 2006; Ju, 2000; Oshita, 2000; Park & Laskhmanan, 2007; Yip, 1994) have recorded similar overgeneration involving unaccusative verbs. Contrary to these studies, overgeneration which is realised in *BE + Ved/en* structure in the L1-Malay data occurs mainly with active verbs: transitive and unergative verbs. Transitive verbs allow for passivisation since they are agentive in nature and they take subjects that assume the role of agents. This suggests that overgeneration involving inflected transitive verbs is not the result of confusion of the

lexico-semantic of the main verbs, but could be attempts learners make to mark tense feature. This possibility is supported by the qualitative findings that reveal instances of the main verbs inflected with 3rd person singular *–s* such as '*It* ***is depends****..'* and '*Learning process in university* ***is begins****..*', which indicate the use of affixal inflection to check tense feature. The consistent use of verbal inflections in marking the present tense in these overgeneration constructions provides strong evidence that learners are utilising the affixal inflections to check the tense feature. The findings suggest that the form of post-*BE* verbs is perhaps determined by learners' attempts to realise the inflectional projection (IP) in their grammar; when *BE* is utilised to mark agreement, the lexical verbs will remain in their infinitival forms, however, they would be inflected when the tense feature needs to be checked.

3. Finite *BE* forms are overgenerated more frequently than non-finite *BE* forms

Overgeneration instances in the L1-Malay learner data involve mainly finite *BE* and only few instances of overgenerated non-finite *BE* forms are recorded, involving mainly infinitive *be* and *been*. Infinitive *be* is overgenerated in *modal + be + V* structure, which produces constructions such as "…*they* <u>*should* ***be*** *collect*</u> *many of money*…", while *been* overgeneration is realised in *have + been + Ved* structure producing constructions such as "…*countries in the world* <u>*have* ***been*** *started*</u> *to join the campaign…*". Both structures suggest misapplication of grammar rules, therefore, can be categorised as intralingual errors. Infinitive *be* overgeneration seems to be the result of overgeneralisation of simple future tense (*modal + be*) as in "*I* ***will be*** *there soon"* or misapplication of passive rule (*modal + be + PP*) as in "*He* ***could be done*** *with that earlier".* Overgeneration involving *been* points to either misapplication of perfect passive rule (*have + been + PP*) as in "*have been called"* and "*had been thrown"* or learners' misconception that auxiliary *have* in present perfect tense needs to also be preceded by *been*.

It is important to highlight that instances of overgeneration involving both infinitive *be* and *been* are too marginal to be considered as serious or persistent. As already mentioned, non-finite forms do not carry tense and agreement as these features are already realised by the auxiliaries or modals preceding them. As a result, non-finite forms do not have to undergo morphological transformation like the finite forms, thus, eliminates the confusion learners' face in realising the tense and agreement features. This has resulted in the relatively low instances of ungrammatical use of non-finite *BE* in this study.

6.2.2.2 Omission of *BE*

Omission of *BE* is the second most frequent type of ungrammatical use in this study. The overall percentage of omission in the data is comparatively low (about 23%) suggesting that omission is not as serious as *BE* overgeneration. Nonetheless, the persistent occurrences of omission in the writing of less proficient learners as revealed by the qualitative analysis indicate that serious attention should be given to this aspect, so that it can be addressed more effectively.

This study has analysed several variables in relation to omission which include the grammatical function of *BE* (copula *BE* and auxiliary *BE*) and syntactic environments (subjects, subject predicates and presence of auxiliaries and intensifiers). The major findings of the patterns of omissions are recapitulated below with a brief discussion and interpretation of each pattern.

1. Higher omission of auxiliary *BE* than copula *BE*

In general the study records higher instances of auxiliary *BE* omission (*Nowadays, we ∅* ***using*** *a lot of money to get make sure the best living we have*.) than copula *BE* (*Co-riricular activities ∅ also important to the student for get a job*.) consistent with the findings of Dulay et al. (1982), Lakshmanan (1995), Milton (2001) Haznedar (2001,

2003) and Ionin and Wexler (2001). Studies analysing the functional categories among child learners acquiring English as L2 (Haznedar, 2001; Ionin & Wexler, 2001; Lakshmanan, 1995) attributed earlier and more consistent supply of copula *BE* than auxiliary *BE* to the universality of L2 language acquisition and that learners regardless of their L1s would follow an almost similar order in the acquisition of English functional categories, similar to the order attested in L1 child acquisition where copula *BE* is acquired before auxiliary *BE* (Brown, 1973; Dulay et al., 1982). Adult L2 learners may have fossilised the omitted structure in their interlanguage as *BE* omission is also a common feature in adult learner language (see Herat, 2005; Muneera & Wong, 2011; Murad & Khalil, 2015).

Several other plausible explanations to account for higher frequency of auxiliary *BE* include, semantic emptiness of *BE*, heavier processing load of auxiliary *BE* and complexity of auxiliary *BE* construction. According to Kuzraj (1985), copula *BE* is semantically more meaningful than auxiliary *BE* and has a number of different meanings compared to auxiliary *BE*, making it easier for learners to relate to, hence use correctly. The notion of emptiness of *BE* is a possible factor instigating omission, however, it is difficult to prove in the context of the current study as the L1-Malay learners are also observed to omit quite a number of copula *BE*.

Previous research also suggests that omission may result from heavy processing load and learners' limited processing capacities. Becker (2004) suggested that the processing load depends on the length of an utterance, the longer it is, the heavier the processing load. Generally, copular constructions tend to be shorter than progressive and passive constructions. However, it is difficult to determine if the omission instances in the L1-Malay data are the result of this factor, as both copula *BE* and auxiliary *BE* constructions are generally found in longer and structurally more complex clauses.

Finally, the complexity of auxiliary *BE* constructions has also been accounted for its higher omission instances. Auxiliary *BE* construction consists of another element in the verb phrase i.e. a main verb (e.g. *is going, are called*) and it is realised in longer construction compared to copula *BE* construction in which *BE* is the only verb. The complexity of auxiliary *BE* clauses is a plausible explanation for higher auxiliary *BE* omissions as learners are also required to consider the inflectional properties of *BE* when constructing auxiliary *BE* clauses.

2. Subject constraint in omission of *BE*

Both copula *BE* and auxiliary *BE* omissions occur more frequently after NP subjects (i.e. 59% in copula *BE*, 65% in auxiliary *BE*) than after PN subjects, consistent with the findings of Ellis (1988) and Herat (2005), who also reported higher *BE* absence after NP subjects. According to Tode (2003, 2007), this is the result of the difficulty learners face in determining the number of the subject and to assign the correct morphological form of *BE*. The supply of *BE* after pronouns subjects is relatively easier as pronouns would not require any suffixation. There is also the tendency for pronoun subjects to be supplied as formulaic sequences as specific pronoun will always be sequenced with specific morphological form for example *it is..., they are..., you are...* (Wilson, 2003). The textual analysis conducted on *BE* omissions in the L1-Malay learner data also reveals that they are commonly preceded by plural noun subjects, confirming the supposition forwarded by Tode (2003, 2007) that L2 learners' supply of *BE* is constrained by number of the subject NPs.

3. Subject predicate constraint in copula *BE* omission

Null *BE* occurs mostly before AP (60%) and secondly before NP (26%) in the L1-Malay learner sub-corpus consistent with the finding of previous studies (Herat, 2005; Lee & Huang, 2004; Unlu & Hatipoglu, 2012). This finding suggests that L1-Malay learners tend to be constrained mostly by *AP* predicates. According to Herat (2005)

omission before adjectival bears similar pattern to omission before past participle environment (*Ved*). Using Holm's (1984) argument, Herat explained that adjectives behave similarly to *–ed* verb. Since adjective can function as a sub-class of verb, thus, would also be favourable to *BE* absence similar to *-ed* verb (p. 199). *BE* omissions in the L1-Malay learner data, however, could not be explained using this principle as they appear to favour adjectivals but not *–ed* verb.

Lee and Huang (2004), however, attributed higher frequency of omission in *BE*-adjective structure to interlingual transfer claiming that copula-like *shi* in Mandarin is used in *BE*-noun structure but not in *BE*-adjective and *BE*-preposition structures. Similarly, past studies on Malaysian learner data attributed omissions in the Malay learner data to transfer from the Malay grammar (Maros et al., 2007; Wee, 2009; Wee, Sim & Kamaruzam, 2010). In determining if interlingual transfer contributes to omission of *BE* in this study, the aspect is dealt with in detail in (4).

4. Interlingual transfer and omission of *BE*

In Malay the use of copula-like verbs '*ialah'* and '*adalah'* are only restricted to *NP + NP* structure when both the subject and its predicate have the same meaning. In addition, they are also used in the *NP + PP* and *NP + AP* structures in the condition when the subject is linked to a predicate that describes or qualifies the subject (Nik Safiah et al., 2010). It is important to note that '*ialah'* and '*adalah'* are not the equivalent of copula *BE* and they are not always required in the Malay grammar even in the *NP + AP*, *NP + NP* and *NP + PP* structures. For instance, sentences such as "Mr Lim is very nice", "He is a teacher" and "He is from Penang" can be expressed in Malay without the use of either '*ialah'* or '*adalah'*:

(1) a. *Mr. Lim sangat baik.* (*NP + AP*)
Mr. Lim ∅ very kind.

b. *Dia seorang cikgu* (*NP + NP*)

He ∅ a teacher

c. *Dia dari Penang.* (*NP + PP*)
He ∅ from Penang.

Considering that Malay has no copula-like verb, it seems obvious that omissions could be the result of negative interlingual transfer from the Malay grammar. If this is the case, the percentages of occurrence for omissions before adjective and noun predicates should be almost similar. In Malay both structures can be expressed without the use of copula *BE*, and if interlingual transfer is the cause of omissions, L1-Malay learners are then expected to omit *BE* at almost similar percentages regardless of the types of the subject predicates. However, in this study there is a vast difference in the percentage of omissions before adjectival predicates and before nominal predicates; 60% and 26% respectively.

Unlu and Hatipoglu (2012) had also ruled out L1 transfer in the variability in the use of *BE* by Russian ESL learners. In Russian copula *BE* is omitted as a rule in present simple tense (PST), resulting in PST constructions such as *I-salesclerk*, *My mother-teacher* and *He-interesting* (p. 256). The researchers found very minimal occurrences of *BE* omission, instead learners were found to substitute *BE* with auxiliary verb *do/does*. According to Unlu & Hatipoglu (2012), it was the outcome of incomplete understanding and application of the rule of copula *BE*, therefore, not the result of transfer from Russian grammar.

Similarly, Herat (2005) came to the same conclusion concerning *BE* absence found in spoken Sri Lankan English, which was found to be consistent with that of other New English varieties (e.g. Singaporean & Malaysian English) and creoles (e.g. South African Indian English). On that basis Herat (2005) argued that zero copula in Sri Lankan English could not simply be attributed to L1 transfer. Instead the researcher explained that it may result from universal grammar tendencies (p. 206). Based on the

findings of this study and that of previous research (Herat, 2005; Unlu & Hatipoglu, 2012) interlingual transfer could not provide convincing explanation for *BE* omissions in this study.

5. The presence of intensifiers and omission of *BE*

Omissions of copula *BE* complemented by AP and NP predicates tend to also occur in the presence of negation *not* and degree adverbs such as *always, very, so,* or *only*. This tendency was also recorded among Chinese learners (Lee & Huang, 2004) and Singaporean English speakers (Ho & Platt, 1993), suggesting that omission of *BE* could be sensitive to the presence of degree adverbs and negation *not*. Nevertheless, compared to negation *not,* omissions before degree adverbs are found to be more frequent in the L1-Malay learner data, which suggests that the presence of degree adverbs exerts slightly more influence over copular omissions than negation *not*.

6.2.2.2 Summary of the Patterns of Ungrammatical Use of *BE*

Overgeneration of *BE* in the L1-Malay learner data which involves mainly *BE* inserted before uninflected transitive verbs supported the findings of Ionin and Wexler (2001), who postulated that *BE* in *BE + bare V* is a default for instantiating agreement. The system inherent in overgeneration instances in the L1-Malay learner data also suggests that they are intralingual errors, which have rooted from faulty generalisation of the rules of the English grammar.

The patterns of omission of *BE* in this study indicate that omission is constraint by two variables, namely the type of subjects and subjects predicates. Noun subjects appear to instigate more omissions than pronoun subjects. Past researchers (Tode, 2003, 2007; Wilson, 2003) attributed this pattern to the difficulty learners experience in assessing the surface morphology of *BE*. The qualitative findings of the current study corroborate this supposition as noun subjects are found to trigger more *BE* omissions than pronoun

subjects. As for omissions that favour adjective predicates, they are most often associated with negative interlingual transfer (Lee & Huang, 2004; Maros et al., 2007; Wee, 2009; Wee, Sim & Kamaruzam, 2010). Nevertheless, this study rejects this notion as interlingual transfer could not adequately explain the variability in the percentage of omissions in *BE*-adjective and *BE*-noun sequences, when in the Malay grammar both can be expressed without the use copula-like verb.

6.3 Influence of Syntactic Environments on Grammatical and Ungrammatical Uses of *BE*

The discussion in this section focuses mainly on the role of the syntactic environments on the grammatical and ungrammatical uses of *BE*. It aims to address the fourth research question of this study:

> RQ4: How do the syntactic environments influence the grammatical and ungrammatical uses of *BE* in the essays written by L1-Malay ESL learners?

6.3.1 Influence of Syntactic Environments on Grammatical Use of *BE*

The syntactic environments under investigation for the grammatical use of *BE* include the types of subjects, subject predicates and the class of post-*BE* lexical verbs. The influence of each constituent is discussed separately in the subsequent sections.

6.3.1.1 Types of Subjects

NP subjects appear to be the most frequently used compared to the other types of subjects in all the major functions; 64% in copula *BE* constructions and 66% in auxiliary *BE* constructions. One factor that could contribute to the high frequencies of NP subjects is the repeated use of the key nouns taken from the essays prompts. Learners were given titles such "*Crime does not pay*" and "*Money is the root of all evil*", and they tend to constantly use the key nouns for example "*crime*" and "*money*"

in their essays. The repetition of the key nouns may have resulted in the high frequency of NP subjects in the L1-Malay learner sub-corpus. This tendency is common in essay writing as learners need to make constant reference to the subjects of the essays in their strategy to draw focus to the issue being discussed (Biber et al., 1999).

Personal pronouns (PPN) are found to be the second most used subjects in the *BE* constructions compared to other pronoun categories such as demonstrative and indefinite pronouns. PPN generally serves as references to replace noun phrases in texts. They are better suited for that purpose since they can be very specific and cover a wide range of subjects/items. There are different forms of PPN that are categorised according to number, person, case and gender making them better choices as references.

Some PPN such as *we*, *you*, and *they* have generic usages. In writing, it is typical for these pronouns to be used to refer to people in general. The use of these pronouns also evokes a sense of commonness, appeals to common human experience and invites empathy from the readers (Biber et al., 1999). There are possibilities of the learners use them as a strategy to appeal to their readers. Therefore, it is not surprising for personal pronouns such as *we* and *they* to be frequently used in the learner essays. Nevertheless, it is very difficult to determine the extent of the influence that the types of subjects have on the grammatical use of *BE*. It is clear that at the ultimate attainment stage, mapping of subjects to the appropriate verbal morphology is no longer a challenge to the majority of the learners in this study.

6.3.1.2 Types of Subject Predicates

Consistent with the patterns of copula *BE* constructions in academic prose written by native speaker writers in LSWEC (Biber et al., 1999), copular constructions in the L1-Malay learner data constitute mostly of *BE*-noun and *BE*-adjective structures. Both structures are consistently used in declaratives, negatives and interrogatives. The

findings from both quantitative and qualitative analyses reveal that most learners have acquired these copular structures well and possess the competency to use them correctly.

As mentioned earlier, *BE*-noun and *BE*-adjective structures perform specific functions in academic writing, resulting in them being employed more frequently in such texts compared to other *BE*-copula structures. According to Biber et al. (1999) noun predicate is commonly used to either characterise a subject or to identify the subject of the noun phrase, while adjective predicates are most commonly used in academic prose to express intellectual claims. Normally the adjectives would be complemented by a prepositional phrase, *that*-clause or infinitive *to* clause, in which the claims are expressed (Biber et al., 1999). Nevertheless, it is difficult to determine the influence of subject predicates in the grammatical constructions of *BE*-copula in this study. Learners are found to be proficient in all the major *BE*-copula sequences, but, tend to favour *BE*-noun and *BE*-adjective following the requirement of academic writing.

6.3.1.3 Class of Post-*BE* Verbs

The use of auxiliary *BE* in the formation of passive voice and in the progressive aspect is analysed in relation to the post-*BE* verbs, namely transitive verbs (Vt), unergative verbs (Uer) and unaccusative verbs (Uac). Transitive verbs are found to be used most frequently in the auxiliary progressive constructions and the only class of verb used in the formation of passives. The finding suggests that learners are well aware that passive voice can only be formed with transitive verbs, unlike progressive constructions which allow for the use of both transitive and intransitive verbs. This is a manifestation of learners' competency in the deep structure of both transitive and intransitive verbs. In addition, at this stage learners also display advanced competency in the intricate process of subject-object inversion that is primary in the construction of passives. The class of

post-*BE* verbs, however, could not be directly associated with the grammatical construction of passives and progressives in the L1-Malay learner sub-corpus. However, it can be concluded that the L1-Malay learners in this study show high level of competency in the use of auxiliary *BE* in both passive and progressive constructions.

6.3.1.4 Summary of the Influence of Syntactic Environments on Grammatical Use of *BE*

As stressed earlier, the syntactic environments could not be directly associated with the grammatical use of *BE* in this study. Nevertheless, the analysis of the constituents surrounding *BE* has managed to provide a comprehensive account of the patterns of the grammatical constructions of *BE*.

Copula *BE* constructions in this study are most frequently preceded by NP or PPN subjects and complemented by either NP or AP predicates. This does not necessarily mean that the learners are unable to produce grammatically correct copula *BE* constructions with other types of subjects or subject predicatives. The frequent occurrences of NP and PPN as subjects and NP and AP as subject predicates are consistent with their occurrences in the native speaker corpus (Biber et al., 1999). This shows that L1-Malay learners are using these items in the same pattern as the native speakers, which suggests that the frequency of use cannot be linked to learners' proficiency. Some structures occur less frequently due to perhaps the learners' writing style, the writing genre or the essay topics, thus, not the outcome of learners' incompetency in the language.

As for auxiliary *BE*, both auxiliary passive and auxiliary progressive constructions favour NP and PPN as the subjects. In terms of the post-*BE* verbs, transitive verbs tend to be used most frequently in the auxiliary *BE* progressive constructions, while passive constructions occur exclusively with transitive verbs. Despite the higher frequency of

transitive verbs in the auxiliary *BE* constructions, the findings have also proved that the learners are able to use intransitive verbs correctly. This is evident in the correct use of unergative and unaccusative verbs in the auxiliary progressive constructions.

6.3.2 Influence of Syntactic Environments on Ungrammatical Use of *BE*

This section discusses the influence of the syntactic environments on the ungrammatical use of *BE*. The syntactic environments examined include the types of subjects, subject predicates, class of post-null *BE* verbs and the presence of intensifiers and modal auxiliaries.

6.3.2.1 Types of Subjects

The analysis of the influence of the types of subjects in the major ungrammatical use of *BE*; overgeneration and omission, shows mixed results. Types of subjects are found to have no influence over overgeneration of *BE*, but seem to exert some influence over *BE* omission.

In the instances of overgeneration *BE* tends to be overgenerated slightly more often after pronoun (54%) than noun (46%) subjects. The relatively small difference in the percentages suggests an unlikely influence of the type of subjects on the overgeneration instances. The pattern of *BE* omission, however, suggests otherwise. Omissions of *BE* are more prominent after nominal compared to pronominal subjects. This tendency is believed to be the outcome of the underlying process in the acquisition of English inflectional projection (IP) system. The system is argued to be acquired in 'chunks', which are heavily rooted in lexically specific constructions such as *He's/It's/I'm* (Pine et al., 2008; Rice et al., 1998; Wilson, 2003). *BE* is acquired in the similar way a lexicon is acquired, in the subject-*BE* combinations (*PN + BE*) such as *they are, it is, you are* or *he is,* which are very often reduced to contractions *they're, it's, you're* and *he's.* The same contraction mechanism, however, could not be applied to noun-*BE* sequence. The

supply of *BE* after a noun subject would require for the learners to determine the number of the subject before deciding the morphological form of *BE*, which according to Tode (2003, 2007) could cause confusion to L2 learners. The supply of *BE* after pronouns subjects is also relatively easier as pronouns would not require any suffixation.

The findings on the types of subjects in this study suggest a strong influence of the types of subjects over *BE* omissions. This supposition is supported by Ellis (1988) and Herat (2005), who also reported consistent supply of *BE* after pronominal subjects than after nominal subjects. Based on the findings of this study and that of previous research, it can be concluded that types of subjects could have some influence on the omission of *BE* by the L1-Malay learners.

6.3.2.2 Class of Post-*BE* Verbs

The class of post-*BE* verbs also has little influence over all the major ungrammatical use of *BE* in the L1-Malay learner sub-corpus. Overgeneration and omission errors are found to involve mainly transitive verbs, which is not surprising as transitive verbs are used more often in the learner data compared to intransitive verbs (i.e. unergative and unaccusative verbs).

Past studies have recorded the tendency for overgeneration to occur with unaccusative verbs (Hirakawa, 2006; Ju, 2000; Oshita, 2000; Park & Laskhmanan 2007; Yip, 1994), which Oshita (2000) termed as overpassivisation due to its resemblance to passive construction. Overgeneration instances in the L1-Malay learner data do not occur exclusively with unaccusative verbs, instead they occur more often with transitive verbs and are believed to be triggered by the need to instantiate agreement. Suppletive inflections (*BE*) is inserted before a lexical verb to perform this task. An important

feature of overgeneration is the lexical verb would be left bare taking the *BE + V* structure, in contrast overpassivisation is realised in *BE + Ved/en* sequence.

The same pattern is also observed in *BE* omissions, which mainly involve transitive verbs. As mentioned earlier, transitive verbs are used more frequently than intransitive verbs in the L1-Malay learner data, thus, it is not unusual for omissions to occur more frequently with transitive verbs. The pattern of omission of *BE* with regard to the class of verbs shows that omission is not determined by the class of post-*BE* verbs. It can be concluded that class of post-*BE* verbs has no bearing on the major types of ungrammatical use of *BE* among the L1-Malay learners in this study.

6.3.2.3 Types of Subject Predicates

In this study, the influence of subject predicates only concerns the omissions of copula *BE*. The quantitative analysis reveals more instances of *BE* being omitted before adjectival predicates compared to before nominal predicates. The finding suggests that learners may find *BE-adjective* more difficult than *BE-noun* sequence. Similarly, Sri Lankan English speakers (Herat, 2005), Chinese (Lee & Huang, 2004), and Russian (Unlu & Hatipoglu, 2012) learners were also found to favour zero *BE* before adjective predicates. These findings suggest possible influence of type of predicates on *BE* omissions in this study and at the same time provide support that variable use of *BE* could be the result of developmental tendencies as postulated by Herat (2005) and Unlu and Hatipoglu (2012).

6.3.2.4 The Presence of Intensifiers and Modal Auxiliaries

Overgeneration and omission of *BE* in this study occur more frequently without the presence of either intensifiers or modal auxiliaries. Even though copula *BE* omissions especially those complemented by adjectivals tend to occur in the presence of degree adverbs such as *always*, *so* or *very*, their occurrences are still relatively fewer compared

to those without the presence of the degree adverbs. Overgeneration and omission of *BE* also tend to occur without the presence of modal auxiliaries. Based on these findings, it is clear that the presence of intensifiers and modal auxiliaries are not strong influences to both *BE* overgeneration and *BE* omission in this study.

6.3.2.5 Summary of the Influence of Syntactic Environments on Ungrammatical Use of *BE*

The analysis of the syntactic environments in overgeneration and omission of *BE* reveals that only the types of subjects and subject predicates appear to influence the ungrammatical use of *BE* in particular omission of *BE*. As discussed earlier, the influence is believed to stem from the developmental aspect of language acquisition. Omissions, which tend to occur more frequently after noun subjects, is closely linked to how learners acquire the English IP system, which has resulted in a more consistent and accurate supply of *BE* after pronominal than after nominal subjects (Tode, 2003, 2007; Wilson, 2003).

The same explanation could also be offered for *BE* overgeneration in the L1-Malay learner data. The patterns of overgeneration in the learner data suggest that overgeneration is also the outcome of the projection of the English IP system. This can be traced through the function performed by *BE,* that is as an agreement marker. This overgeneration structure is consistently used by the learners with both transitive and intransitive verbs.

The similarity in the variability in the supply of *BE* by the L1-Malay learners with other L2 learners also provide further evidence that the errors are the result of the developmental aspect of language acquisition. L1-Malay learners tend to be constrained by *BE-adjective*. The same tendency is also evident in the language data of English speakers and learners from other L1 backgrounds including L1-Chinese (Lee & Huang,

2004), L1-Sinhala (Herat, 2005) and L1-Russian (Unlu & Hatipoglu, 2012), providing evidence that learners regardless of their L1s would encounter similar difficulties with the use of *BE*. Similar to Malay, Sinhala and Russian grammars do not have copula-like verb, which makes *BE* omissions the likely outcome of interlingual transfer. Unlu and Hatipoglu (2012), however, reject this notion as L1 interlingual transfer could not adequately explain why Russian learner replaced *BE* with auxiliary *do/does*. Similarly, Herat (2005) also rejected the notion of interlingual transfer, and argued that the similarities in the variable use of *BE* in spoken Sri Lankan English with other new Englishes or creoles as evidence of developmental aspect of acquisition. On the basis of the findings of the current and previous studies it is concluded that the syntactic environments bear only some influence on the ungrammatical use of *BE* in this study and at the same time confirming the supposition of previous studies (Herat, 2005; Unlu & Hatipoglu, 2012) that the variable supply of *BE* is "the result of universal grammar tendencies" (Herat, 2005, p. 206).

CHAPTER 7

CONCLUSIONS

7.0 Introduction

This chapter concludes the study by giving a comprehensive account of the use of all forms and functions of *BE* in the L1-Malay learner data. It identifies the patterns of the grammatical and ungrammatical uses of *BE* and the syntactic environments that may influence the patterns of use. It summarises the main findings, discusses its theoretical, methodological and pedagogical implications and highlights some of its limitations.

7.1 Main Findings

This section is divided into two; the first addresses the analytical aspect of the study by summarising the overall distribution of *BE*, the patterns of the grammatical and ungrammatical uses of *BE* and the influence of the syntactic environments on the patterns of use, while the second addresses the application of the findings of the study by briefly summarising the corpus consultation model proposed for the teaching of *BE* to ESL learners in Malaysian universities.

7.1.1 Main Findings on the Use of *BE*

7.1.1.1 Overall Distribution of *BE*

The corpus-based methodology used in this study enables the researcher to reveal both the grammatical and ungrammatical uses of *BE*, which are lacking in past studies investigating *BE*. Previous research focuses mainly on learner errors on *BE*, thus, judgment on learners' competency in the use of *BE* derived solely on what learners could not perform with the verb, without any regard to what they could. This study goes

beyond learner errors and instead promote a concurrent analysis of the correct and incorrect uses of *BE*. By focusing on both aspects, this study is able to provide different insights into the use of *BE* by the L1-Malay ESL learners.

One of the major findings in the overall distribution of *BE* is the significantly higher percentage of the grammatical use of all *BE* forms and functions in the L1-Malay learner data. In addition, the qualitative analysis reveals that *BE* is most commonly used in structurally complex constructions. They are used in extended main or subordinate clauses in various constructions, namely declarative, interrogative, negative, existential *there* and *it*-cleft. These findings are the opposite of the findings from previous error analysis research that reported prevalent errors of *BE* at different stages of language learning among L1-Malay learners in Malaysia (Arshad & Hawanum, 2010; Jishvithaa et al., 2013; Manokaran et al., 2013; Maros et al., 2007; Ting et al., 2010; Wee, 2009; Wee, Sim & Kamaruzam, 2010).

The comparative analysis conducted between L1-Malay sub-corpus and native learner sub-corpora (LOCNESS) has also led to an important conclusion that the Malay and American learners share similar patterns in the use of *BE*. The study also reveals that learners' grammatical use of *BE* is strongly influenced by the writing genre. There is a heavy reliance on *BE*-copula constructions and passives, the two types of constructions ascertained to be more prominent in academic prose (Biber et al., 1999; Hinkel, 2002; Swales & Feak, 2012). There is also a more frequent use of the present tense in the L1-Malay learner data, which is another common feature of academic prose. According to Biber et al. (1999) and Hunston (2002) simple present tense is widely used in academic prose to convey general truth. It is also used to express a wider range of meanings referring to past events, present states, habitual behaviours and future events (Biber et al., 1999), making it more common in academic writing.

The study has also revealed persistent errors in the use of *BE*. Even though errors in the use of *BE* are not widespread, two are found to be very persistent, namely *BE* overgeneration and *BE* omission. Overgeneration stems from the need to instantiate agreement. Instead of applying affixal inflection to a main verb, *BE* is inserted before the main verb to mark agreement, producing deviant verbal clauses such as "*is take, are study"*. Overgeneration in the L1-Malay learner data most frequently involves uninflected active verbs (transitive and unergative). Textual analysis reveals consistent subject-verb concord in the instances of overgeneration, which provides further support that overgenerated *BE* functions as the marker for agreement. The system inherent in the overgeneration instances suggest that they are intralingual errors as they possess characteristics of rule learning (Richards, 1971) and that the system has been fossilised in the learners' interlanguage (Selinker, 1972).

As for omissions, the findings reveal that auxiliary *BE* tends to be omitted more frequently than copula *BE*. *BE*-auxiliary is structurally more complex compared to *BE*-copula sequence, as the verb phrase requires an additional element in the form of a main verb. This structural complexity is a possible cause for more frequent instances of auxiliary *BE* omissions. Both auxiliary *BE* and copula *BE* omissions tend to be constrained by noun subjects, which suggests that mapping noun subjects to the correct morphological forms of *BE* is still a challenge to some learners. Similar to overgenerations, omissions also bear characteristics of interlanguage fossilisation. The textual analysis also reveals that *BE* overgenerations and *BE* omissions on their own account do not hamper communication, but when combined with other language impairments (e.g. wrong lexical choice or sentence structure) they can affect communication and reduce the overall quality of learner compositions.

The qualitative findings also highlight several differences in the use of *BE* between proficient and less proficient learners. Proficient learners tend to use fewer *BE* in their

writings and most often *BE* is used in structurally complex constructions. Whereas, less proficient learners tend to use *BE* more frequently and most often the verb is used incorrectly. The ungrammatical use of *BE* strongly suggests that less proficient learners are still grappling with elementary issues such as agreement and verbal morphology. This has very important pedagogical implications as teachers not only need to focus their attention on the efforts to address the specific problems on *BE*, but at the same time tackle other language issues, so that the gap between proficient and less proficient learners can be reduced.

7.1.1.2 Patterns of Grammatical and Ungrammatical Uses of *BE*

The corpus-based approach employed in this study has also allowed very comprehensive analyses of *BE* to be conducted. They include not only the analyses of the occurrences of each form and function of *BE*, but also the constituents surrounding the verb. As a result, this study has successfully revealed the patterns of the grammatical and ungrammatical uses of all the finite and non-finite *BE* forms.

7.1.1.2.1 Patterns of Grammatical Use

L1-Malay learners are able to use correctly all the major functions of finite and non-finite *BE* forms, which include the use of (i) finite *BE* in the copula and auxiliary *BE* in the construction of declaratives, interrogatives and negatives (ii) non-finite *BE* in the formation of simple future tense, future passive, future progressive, present/past perfect, perfect passive, perfect progressive and progressive passive, and (iii) *BE* in the construction of existential *there* and *it*-clefts.

Some constructions occur more frequently than others due to several factors, which include the register (written), writing genre and essay prompts. Below is the summary of the major patterns of the finite (1-3) and non-finite *BE* (4-6) in the L1-Malay learner sub-corpus.

1. Copula *BE* constructions in the L1-Malay learner data are often preceded by either NP or PPN subjects and followed mainly by AP and NP predicates as shown in (a) and (b):

   (a) *NP/PPN + BE + AP*

   (b) *NP/PPN + BE + NP*

2. Auxiliary *BE* in passive voice is most frequently preceded by either an NP or PPN subject and followed by the past participle of a transitive verb as shown in (c) and (d) below:

   (c) *NP/PPN + BE + Vt-ed + PP*

   (d) *NP/PPN + BE + Vt-ed + by-phrase*

3. Auxiliary *BE* in progressive aspect also tends to be preceded by NP or PPN subject and followed by the present participle of a transitive verb as shown in (e) below:

   (e) *NP/PPN + BE + Vt-ing + expansion*

Even though non-finite *BE* forms are used less frequently than the finite forms, they tend to be used very consistently. They are most often used to perform various functions realised in specific sequences summarised below:

4. Infinitive *be* is used most frequently in the formation of simple future tense (*modal + be*) and passives (*modal + be + Ved*). The simple future constructions are often preceded by NP or PPN subjects and complemented by either AP or NP predicates as in (a), while the passive constructions are most often followed by transitive verbs as in (b) as shown below:

   (a) *NP/PPN + modal + be + AP/NP*

   (b) *NP/PPN + modal + be + Vt-ed + expansion*

5. *Been* is used mainly in the formation of perfect passive (*have/has/had + been + Ved*). These constructions are frequently preceded by NP or PPN subjects and

most often followed by transitive verbs. The construction of *been* is as shown in (c):

(c) *NP/PPN + have + been + Vt-ed +expansion*

6. *Being* is mainly used in the formation of progressive passive (*BE + being + Ved*). The construction is very rare in the L1-Malay learner sub-corpus. It is often preceded by NP subjects and followed by a transitive verb as shown in (d):

(d) *NP + BE + being + Vt-ed + expansion*

7.1.1.2.2 Patterns of Ungrammatical Use

This study has managed to reveal (i) two major types of ungrammatical use of *BE*: overgeneration and omission, (ii) the patterns of each type of ungrammatical use and (iii) the syntactic environments that might influence the ungrammatical use. The major types of ungrammatical use of finite *BE* are summarised in (1) and (2), while (3) presents the summary of the ungrammatical use non-finite *BE*:

1. Overgeneration often occur after PN or NP subjects, they involve mostly uninflected transitive verbs as shown in (a) below:

   (a) *PN/NP +BE +Vt + Complement*

2. Copula *BE* omissions often occur after noun (NP) or pronoun (PN) subject and before adjectival (AP) or nominal (NP) predicate as in (b) and (c) below:

   (b) *NP/PN + Cop Ø + AP*

   (c) *NP/PN + Cop Ø + NP*

   Omissions of auxiliary *BE* occur most frequently after NP subjects and before transitive verbs as shown in (d):

   (d) *NP + Ø + Vt-ing + expansion*

3. The ungrammatical use of non-finite *BE* mainly constitutes overgeneration of infinitive *be* and *been*. As shown in (f), when infinitive *be* is overgenerated in the *modal + be* structure, it is often preceded by NP or PN subject and complemented by AP predicate. If overgeneration occurs in *modal + be + V* structure, it also tends to be preceded by either NP or PN subject and most often followed by uninflected transitive verbs as shown in (g). Overgeneration involving *been* is also found to favour NP or PN subject and most often preceded by auxiliary *have* as shown in (h).

   (f) *NP/PN + modal + be + AP*
   (g) *NP/PN + modal + be + Vt + expansion*
   (h) *NP/PN + have + been + Ved*

7.1.1.3 Influence of Syntactic Environments on the Use of *BE*

This study has also detected possible influence of the syntactic environments on the ungrammatical use of *BE*, particularly on *BE* omission. *BE* tends to be dropped when it is preceded by a plural noun subject or when complemented by an adjective predicate. The deletion of *BE* after noun subjects is strongly associated how English IP system is acquired. Researchers argued (Tode, 2003, 2007; Wilson, 2003) that the supply of *BE* after pronoun subjects is easier as the verb is acquired in subject-*BE* combinations such as *they are*, *he is* or *it is*. However, the same system could not be applied to noun subjects as learners would have to first determine the number of the noun before deciding the morphological form of *BE*.

Lee and Huang (2004) attributed the absence of *BE* in *BE-adjective* sequence to L1 negative interlingual transfer, as Chinese copula-like verb *shi* can be used to link a subject to a noun predicate but not to an adjective predicate. In the Malay grammar copula-like verbs can be used in the *NP + AP* structures, therefore, negative interlingual transfer could not satisfactorily justify *BE* deletion before an adjective predicate in this

study. Instead, this tendency is believed to be linked to the developmental aspect of acquisition as the same pattern is also attested in the data of learners from other first language backgrounds including Sinhala (Herat, 2005) and Russian (Unlu & Hatipoglu, 2012). Other than the conditions discussed above, the syntactic environments do not appear to influence other instances of ungrammatical use. This leads the researcher to conclude that the syntactic environments only affect some aspects of the ungrammatical use of *BE* in this study.

7.1.2 Application of the Research Findings

Based on the research findings, this study proposes a corpus consultation model (CCM) to address the problematic areas of *BE* and improve ESL learners' writing. The model, which is presented and described in detail in Chapter 8, addresses the application aspect of this study. The findings of the grammatical and ungrammatical uses of *BE* serve as the foundation for the development of the CCM. The decision made on the important features of the model; aspects of *BE* to include, type and size of corpus and the teaching and learning activity, derives mainly from the findings of what learners can and cannot do with *BE*.

The proposed CCM model is an example of how research can be linked to application. This study shows that the information on what learners know and do not know can be used to develop an intervention to address problems learners may face. This is an important contribution to future corpus-based research on learner language. This thesis can be used as a reference for future corpus studies that intend to link corpus findings to practice. According to O'Keeffe, McCarthy and Carter (2007), there is a need for a better synergy between corpus linguistics and language teaching since many of the research questions of corpus-based investigations 'arise out of practice' (p. 246).

The proposed CCM can also serve as a guideline for the integration of corpora for the teaching of other aspects of the English language besides *BE*. The same model can be used to teach for instance, phrasal verbs, prepositions, conjunctions, lexical bundles, transitional markers, idiomatic expressions, and etc. The integration of corpora in language teaching a relatively new phenomenon in Malaysia, and the CCM can be used as a reference by novice users of corpora in developing their own corpus-based language teaching materials.

7.2 Implications of This Study

This section presents the theoretical, methodological and pedagogical implications of this study.

7.2.1 Theoretical Implications

Theoretically, this study contributes specifically to the research on the acquisition of *BE* and the acquisition of inflectional projection system by older and more advanced ESL learners. The present study found that *BE* overgeneration is motivated by the need to instantiate agreement and possibly tense feature too. This finding is consistent with the universal tendency that learners tend to use verbal morphology to mark agreement or/and tense as captured by Missing Surface Inflection Hypothesis-MSIH (Haznedar & Schwartz, 1997). The hypothesis states that Tense and Agreement features are present in L2 grammar, but the learners have problems mapping from the abstract features to the corresponding morphological form and tend to resort to suppletive inflections when they are unsure of which affixal inflections to use (Ionin & Wexler, 2001). The behaviour of *BE* overgeneration in this study is consistent with the difficulty in assessing the surface morphology as postulated by MSIH, hence, provides additional support for the hypothesis with the data of *BE*.

The study also contributes to the second language acquisition in general. Several variables have been identified to potentially influence the use of *BE* including the syntactic environments of *BE*, universal learning mechanism, developmental aspects and L1 transfer. These variables may be working in tandem, resulting in the persistent occurrences of the ill-formed constructions, which are believed to have been fossilised in the learners' interlanguage.

7.2.2 Methodological Implications

This study demonstrates how corpus-based investigation of learner corpora, which contain a huge sample of authentic learner language, can contribute significantly to second language acquisition research. The corpus-based approach adopted for the current study has several advantages over non-corpus-based approaches.

Firstly, the corpus-based approach has enabled examination of a large sample of authentic learner language. The size of the L1-Malay learner sub-corpus would be more representative of the English language of the Malay learners. Therefore, the results from this study would provide a better picture of the state of the English language of the chosen learner sample.

Secondly, corpus-based approached adopted enables a comprehensive analysis of the use of all the forms and functions of *BE* to be conducted. The approach enables not only the analysis of the deviant use of *BE*, but also a detailed analysis of the well-formed constructions. This was made possible with the aid of lexical analysis software such as WordSmith Tools (Scott, 2017). The software also enables the surrounding constituents to be concurrently analysed in order to determine the possible influence of the syntactic environments on the use of *BE*.

Finally, comparative interlanguage analysis approach adopted for this study allows for comparison between NNS learner corpus and NS learner corpora to be conducted. The

analysis provides information on the similarities and differences in the patterns of use of all the forms and functions of *BE* in the NS and NNS learner corpora. The information is valuable in determining the specific patterns preferred by the NNS learners and how similar or different these patterns are to that of the NS learners.

7.2.3 Pedagogical Implications

The findings from this study have important implications to the teaching and learning of *BE* to ESL learners in general.

Firstly, the findings of both the ill-formed and well-formed *BE* constructions provide teachers with valuable input on the learners' strengths and weaknesses. This information is especially useful in deciding a suitable approach to adopt and methods to devise in teaching *BE*. In addition, the information on the syntactic variables influencing *BE* constructions also enables teachers to make more informed decision on the syntactic constituents requiring special attention when designing treatments for problematic areas of *BE*. Omissions of *BE* for instance are strongly triggered by plural noun subjects, therefore, treatment for omission should focus specifically on *BE* used with plural noun subjects. The specific treatment saves teachers' time and energy.

Secondly, the findings from the detailed analysis of *BE* overgeneration and omission provide teachers with the knowledge of the underlying system governing these constructions and enrich their understanding of the factors influencing the constructions. This is especially useful in deciding the treatment to be employed for each type of ill forms. For instance, omission of *BE* helps draw teachers understanding on the problems learners may experience with progressive aspect. Therefore, the treatment for omissions of *BE*, should take into consideration the aspectual aspect of *BE*. The understanding that overgenerated *BE* may be the fossilisation of an interlanguage rule (i.e. *BE* as a marker

for agreement), allows teachers to seek specific solutions that would help learners to defossilise the rule.

Finally, the findings also call for the integration of corpora in the teaching and learning of *BE*. Direct incorporation of corpora in the language classrooms could (i) create opportunities for learners to interact with massive amount of authentic language data, (ii) promote learner autonomy (iii) develop learners' cognitive capability through inductive learning and (iv) integrate multiple language skills. As discussed in detail in Chapter 8, a corpus can be utilised as a reference to correct learner errors. Learners would be able to infer the correct patterns of *BE* usage through the concordances. The method besides helping learners improve their writing, also promoting greater learner autonomy as the learning process involves active and conscious participation from the learners in analysing the target language.

7.3 Limitations of This Study

Despite its success in providing a comprehensive account of the use of all the *BE* forms and functions in the L1-Malay learner sub-corpus, this study still suffers from several limitations.

Firstly, although comprehensive the study has only managed to analyse the use of *BE* by learners from one L1 background (i.e. Malay), when MACLE also consists of language data of L1-Chinese and L1-Tamil learners. The corpus-based analysis of the use of *BE* requires for the data to be manually coded and this is a time-consuming process. Due to time constraint, the researcher had to limit her analysis to only the Malay learner language data. The findings from contrastive analysis of the patterns of *BE* usage by learners from different L1 backgrounds would enable the researcher to establish more accurately the influence of L1 in the acquisition of *BE*.

Secondly, the ESL learner corpus used for the study consists of only the written data. So far, there has not been any spoken corpus developed to represent the speech of advanced ESL learners in Malaysia. It is believed that the establishment of learner corpus, which consists of both the written and spoken data, would greatly benefit future corpus-based research as they permit for deeper investigation into the use of *BE* to be conducted such as, how the different registers affect the patterns of *BE* use.

Thirdly, in order to provide a comprehensive account of the learners' use *BE*, the study has taken a path that is rarely ventured by previous studies, that is to conduct concurrent analysis of the grammatical and ungrammatical uses of *BE*. Previous studies focused primarily on the variable supply of *BE* and less focus was given on the correct use of the verb. As a result, the literature on the grammatical use of *BE* are very scarce, making it a challenge to obtain empirical evidence to corroborate the findings of this study.

Finally, findings of the ungrammatical use of *BE* (i.e. overgeneration and omission) strongly suggests that these errors are developmental in nature. Nevertheless, to arrive to such generalisation, more research involving ESL learners from different language backgrounds are needed. Since the present study focuses only on L1-Malay learner data, it is not able to examine the developmental variables more extensively. Thus, unable to attest firmly that learner errors are the manifestation of the natural order of acquisition. With the availability of various learner corpora worldwide such as the International Corpus of Learner English (ICLE) and the International Corpus of English (ICE), which consist of learner data from different first language backgrounds (e.g. Dutch, French, German, and Japan), comparative studies between these learner corpora are highly possible. These studies would be able to identify the potential variables in second language acquisition and more importantly determine more accurately if learners share similar path in acquiring a target language.

In conclusion, the limitations discussed have highlighted areas of the study that could be further developed. It is the hope of the researcher to expand the research in the future to include the areas highlighted here.

CHAPTER 8

CORPUS CONSULTATION MODEL: INTEGRATION OF CORPUS IN THE TEACHING OF *BE*

8.0 Introduction

As direct application of the research findings this study proposes a corpus consultation model (CCM) for the teaching of *BE* to address the problems associated with *BE* as identified in the findings. Relevant findings from the previous analytical chapters will be included in the discussion to show how CCM can be used to support the learning of English, especially with respect to writing.

8.1 Summary of Major Ungrammatical Use of *BE*

This study finds L1-Malay learners produce two major types of errors on *BE* in their essays, namely overgeneration and omission. *BE* overgeneration in the L1-Malay learner data involves mainly uninflected verbs (*BE + V*), where *BE* functions as a marker for agreement. Instead of using affixal inflection to instantiate agreement, learners resort to suppletive inflection. This indicates that learners are facing difficulties with the inflectional projection. There are also instances of *BE + Ved* overgeneration, which suggest that some learners are unable to differentiate the function of *BE* copula from that of *BE* auxiliary in the passive voice.

Omissions of *BE* involves both copula *BE* and auxiliary *BE*. However, the omission of auxiliary *BE* in progressive aspect is more prevalent in the L1-Malay learner data. This is due to *BE*-auxiliary construction being structurally more complex than copula *BE* construction. *BE*-auxiliary verb phrase requires an additional element that is the main verb. The supply of both auxiliary *BE* and copula *BE* omissions also tends to be

constrained by plural noun subjects. This suggests that learners have difficulty to map plural nouns to the correct morphological forms of *BE*. *BE* also tends to dropped more frequently in *BE-adjective* sequence, which researchers believe to be more difficult than other *BE*-copula sequences (Lee & Huang, 2004). The findings from the detailed analysis of *BE* overgeneration and *BE* omission suggest that L1-Malay ESL learners face the following difficulties:

1. formation and functions of copula *BE*,
2. formation and functions of auxiliary *BE* in progressive aspect,
3. formation and functions of auxiliary *BE* in passive voice, and
4. marking of agreement in copula *BE* and auxiliary *BE* constructions

8.2 Corpus Consultation Model

This section presents the corpus consultation model proposed for the teaching of *BE* to L2 undergraduates in Malaysia as a means to improve their writing. The model consists of two important modules: training and teaching. The section also includes the important aspects to be considered by instructors before implementing the actual corpus consultation.

8.2.1 Preliminary Considerations

Prior to the actual corpus consultation, instructors are advised to consider three most important aspects, which include the selection of the grammatical component to be taught, the choice of corpus in terms of function, type and size and the learning activities to be implemented as summarised in Figure 8.1. This section discusses these aspects further.

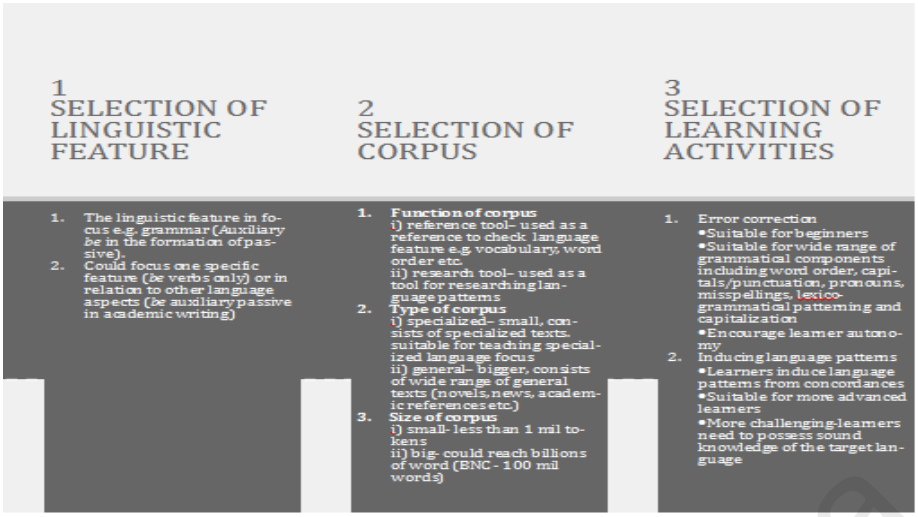


Figure 8.1: Preliminary Considerations for Corpus Consultation

8.2.1.1 Selection of Linguistic Feature

The findings of the ungrammatical use of *BE* reveal several aspects of *BE* that are more difficult to the ESL learners. Firstly, the formation and functions of copula *BE* in the condition when plural nouns are used as the subjects and when *BE* is complemented by adjective predicates. Secondly, the formation and functions of auxiliary *BE* in progressive aspect. Thirdly, the formation and functions of auxiliary *BE* in passive voice. Finally, marking of agreement in copula *BE* and auxiliary *BE* constructions. The confusion and difficulties learners face in these aspects have resulted in *BE* being overgenerated or omitted. The focus of the corpus consultation model is to highlight the ungrammatical use of *BE* (i.e. *BE* overgeneration, *BE* omission) in the learner essays and guide learners to consult the corpus to correct the ill-formed constructions.

8.2.1.2 Selection of Corpus

The section discusses in detail the corpus approach adopted for this study (research tool or reference tool) and the type and size of the corpus selected for the teaching of *BE*.

8.2.1.2.1 The Function of Corpus: Research Tool and Reference Tool

This study proposes for the use of corpus as a reference tool, whereby it is mainly used by the learners as a linguistic reference to solve language and writing problems. Specifically, a tool that learners can utilise to correct errors on *BE* in order to improve their writing. The approach is considered less demanding, as learners are not required to conduct full scale linguistic investigation to deduce language patterns, but are only required to consult the corpus to improve very specific linguistic aspect in their writing (Miceli & Kennedy, 2002; Yoon, 2011). The approach as attested by previous studies (Gaskell & Cobb, 2004; Kennedy & Miceli, 2001; Leel, 2011; Miceli & Kennedy, 2002; Phoocharoensil, 2012; Sun, 2000; Yoon & Hirvela, 2004; Yoon, 2008; Yunus & Awab, 2012, 2014) is more suitable for beginners to corpus consultation, as learners are not presented with the burden of researching the language (Boulton, 2010, 2011). For these reasons the corpus selected for the teaching *BE* in this study is to be used as a reference tool.

8.2.1.2.2 Type and Size of Corpus

The type of the corpora can be divided to two, namely specialised corpora and general corpora. Specialised corpora are strongly associated with the teaching of Language for Specific Purposes (LSP), where the use of custom-built corpora is motivated by the need to cater to the linguistic requirements and interests of specialised linguistic discourses. In contrast, general corpora are bigger and are commonly used to teach general aspects of language such as phrasal verbs or *BE*. General corpora are better suited for the teaching of more general aspects of language that learners regardless of their majors or disciplines find difficult or problematic (Yoon, 2011). The larger size of general corpora (e.g. BNC-100 million words, Collins COBUILD-500 million words) adds to the advantage of using them. They are very useful in providing countless samples of language usages from many different registers, disciplines or contexts. The

corpora provide the learners greater opportunities to interact with the authentic language used in the various contexts, disciplines or registers. Some of them can be accessed free of charge online (e.g. British National Corpus, Collins COBUILD and Corpus of Contemporary American English) and most importantly some are already equipped with built-in concordancers, which allow for simple concordancing to be administered (e.g. British National Corpus, Collins COBUILD and International Corpus of Learner English). This is an important criterion to be considered as not all teaching institutions or schools have the resources and budget to acquire licenses for commercially built concordancers such as WordSmith Tools (Scott, 2017).

In view of the advantages of using general corpora, it is proposed that a general corpus is to be used for the teaching of *BE* to the ESL learners in Malaysia. In doing so, factors such as accessibility and suitability have to be considered. The British National Corpus (BNC) Sampler is proposed to be used for the teaching of *BE* as it is easily accessible and most suitable. BNC Sampler was created and compiled by a consortium whose members include Oxford University Press, Longman Group Ltd, Chambers Harrap, Oxford University Computing Services, UCREL – Lancaster University, and British Library Research and Development Centre. The Sampler consists of 2 sub-corpora; spoken English and written English, each consisting of roughly one (1) million words. For the purpose of this study only the written sub-corpus is to be used. The texts in the written sub-corpus are collections of books, periodicals and other sources covering a range of domains, which include fiction, science, social science, world affairs, commerce, arts, religion and leisure. These materials reflect and represent a wide cross-section of the current British English (http://ucrel.lancs.ac.uk/bnc2sampler/sampler.htm retrieved in May, 2017).

Firstly, the selection of BNC sampler (written) is made based on its accessibility. The corpus can be accessed online at http://www.lextutor.ca/range/, a web-based data-driven

language learning avenue designed to help learners, teachers and researchers conduct corpus-based learning, teaching and research. The web site has multiple corpus-based learning, teaching and research facilities, which also include a web-based concordancer that allows free access to a range of online corpora including the written BNC Sampler. The concordancer is designed with a user-friendly interface. Like other concordancers, the search parameter such as left/right sort, line width and number of result lines can be set prior to a search. For the purpose of simple concordancing the use of BNC Sampler via Lextutor is deemed the most suitable choice as it is not only accessible online but also equipped with a built-in concordancer.

Secondly, BNC sampler is also deemed most suitable for the purpose of teaching *BE* in this study. BNC sampler (written) unlike its main corpus (BNC) only has approximately one (1) million words, which is considered small. The choice for a smaller general corpora is made based on several reasons; (i) *BE* is a common verb, therefore, could be easily found in even smaller corpus, (ii) a large corpus can be intimidating as it generally consists of a very wide range of texts, which might contain unfamiliar topics, words, phrases and complex structures that could be difficult for the learners (Romer, 2011), (iii) a large corpus would also yield large results and considering that *BE* is a common verb, the results obtained from a large corpus would be too massive for the learners to process, and (iv) the difficulties in using *BE* is a common problem among L2 learners regardless of their study majors or disciplines, hence, a general corpus is more suitable for the teaching and learning of more general language problems (Yoon, 2011). In view of these reasons the choice for a smaller general corpus is most suitable for the teaching of *BE* proposed in this study.

8.2.1.3 Selection of Corpus-Based Learning Activities

It is proposed that self-correction of errors through corpus consultation to be implemented for the teaching of problematic areas of *BE*. The choice is made based on the considerations presented and discussed in the subsequent paragraphs.

Firstly, it has been empirically proven that learners performed well in error correction activities (Chambers & O'Sullivan, 2004; Gaskell & Cobb, 2004; O'Sullivan & Chambers, 2006; Todd, 2001). They were reported to do well in a wide range of grammatical components including word order, capitals/punctuation, pronouns (Gaskell & Cobb, 2004), misspellings, lexico-grammatical patterning and capitalisation (Chambers & O'Sullivan, 2004; O'Sullivan & Chambers, 2006). By self-correcting the errors learners had also improved their general writing skills, thus made them more confident writers (Chambers & O'Sullivan, 2004; Gaskell & Cobb, 2004; Kennedy & Miceli, 2001; Kotamjani, Razavi & Hussin, 2017; Miceli & Kennedy, 2002; O'Sullivan & Chambers, 2006; Yoon, 2008; Yoon & Hirvela, 2004). According to Yoon (2008), the corpus can provide the textual help learners need to improve their writing. As a result learners would become more confident and would regard writing as less burdensome. Corpus consultation can also empower learners as they are given the opportunity to be autonomous in language learning through a simple act of error correction. Error correction also provides interactive feedback, where learners can be actively involved in the development of their own language skills. This makes error correction an important and a positive stage in the language learning process (O'Sullivan, 2007).

Secondly, corpus-based error correction activity has proven to be an effective method to overcome word and sentence-level errors and improve learners' general writing skills (Chambers & O'Sullivan, 2004; Gaskell & Cobb, 2004; Kennedy & Miceli, 2001;

Miceli & Kennedy, 2002; O'Sullivan & Chambers, 2006; Todd, 2001). By self-correcting their errors learners are made aware of the language patterns. They self-discover language rules and learn the language in the process of self-discovering (Makino, 1993). Self-correction is considered an important process in language learning and it is "viewed as a global goal of language learning" (Todd, 2001). One of the goals of language learning is for the learners to be able to initiate self-repair (Allwright & Bailey, 1991). In short, correcting errors provide opportunities for learners to acquire the target language as Gaskell and Cobb (2004) put it "an error on a page is an important opportunity in acquisition" (p. 304).

Thirdly, *BE* has a fixed number of inflections and there are very clear formation structures of both copula and auxiliary *BE* constructions i.e. a copular is always followed by a complement and an auxiliary *BE* for marking progressive aspect would be preceded with *Ving*. It is anticipated that learners would find correcting the structure of *BE* through concordancing as not overly complicated.

Fourthly, working with a corpus can be very demanding and overwhelming. Learners would have to deal with an approach that they are not familiar with, the number of search results that can be intimidating and the computer technology that they might not be familiar with. It is paramount that the learners' initial encounter with the corpora be non-threatening. Self-correction of errors is deemed the least threatening approach to corpus consultation.

Fifthly, self-correction of errors can also be an effective way to promote inductive learning. Corpus-based error correction requires learners to perform an active and conscious role in observing the language pattern (O'Sullivan, 2007). It involves two major stages of the language learning process, namely analysing language samples and generating language rules or patterns. Before a pattern can be generated, learners need

to carefully select relevant examples and this process according to Schmidt (1990) and van Lier (1996) is a vital prerequisite for learning.

Considering the positive effects of self-correction on learners' performance in the acquisition of specific language structures and in improving general writing skills, the successful implementation of self-error-correction activities in previous research and its nature of being less demanding and threatening, it is anticipated that self-correction of errors would be most useful in helping the ESL learners in Malaysia to overcome problems with the use of *BE,* hence improve their writing.

8.2.2 Corpus Consultation Model

This section discusses in details the suggested CCM in the teaching of *BE* to the ESL undergraduates in Malaysia. The CCM is divided into two phases; Phase 1 consists of the training component and Phase 2 is the actual teaching and learning component.

8.2.2.1 Phase 1: The Training Component

One focal element to be considered to ensure successful integration of corpora is to provide learners' with adequate training in concordancing. Yoon (2011) highlighted that lack of training and guidance to conduct searchers and interpret results and lack of training in operating the computer and concordancing software are among the major factors that contribute to unsuccessful incorporation of corpora.

Learners require gradual and guided training that can accommodate their different learning styles, experience and language proficiency levels (Yoon, 2011). This type of training can reduce the "cognitive burden" (Boulton, 2010) at the initial stage of the corpus integration, thus, allow learners to familiarise themselves with concordancing. Two aspects of corpus consultation that need specific training and guidance are (i) formulation of search terms and (ii) interpretation of search results. Training on the

hardware and software technologies is also required. The lack of proficiency in IT related functions can also be the source of learners' difficulties and frustration (Yoon & Hirvela, 2004).

In view of the needs for proper corpus training to be conducted, this study proposes for a training component to be implemented before learners can begin to formally consult the corpus. Figure 8.2 below summarises the proposed training component.

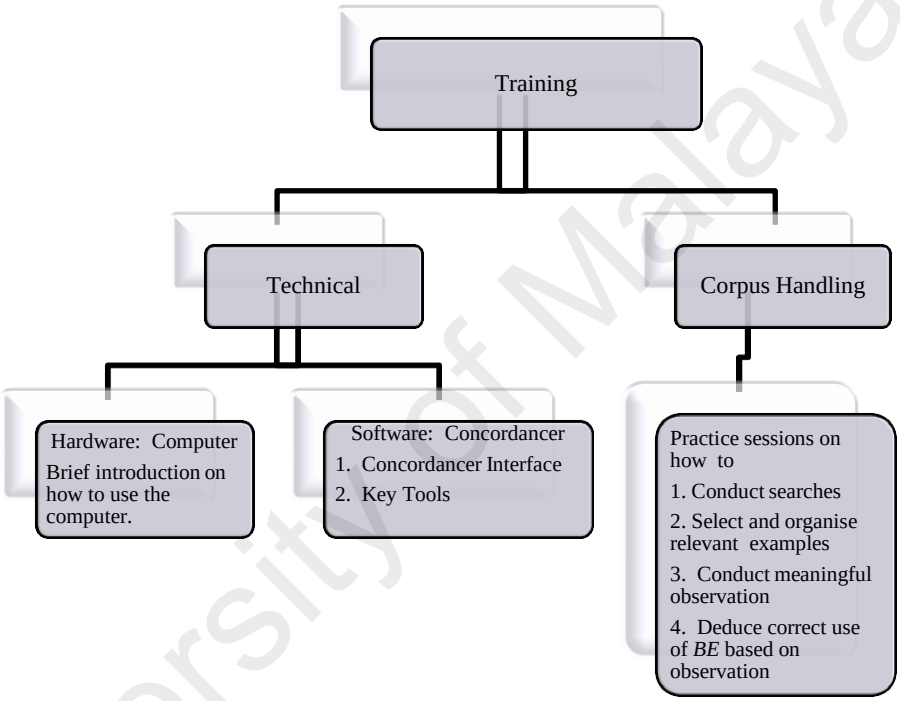


Figure 8.2: Training Component.

As shown in Figure 8.2, the training component is divided into technical and corpus-handling trainings. It is suggested that approximately 8 hours be allocated to the technical training: 2 hours allocated for an introduction to the computer hardware and the remaining 6 hours for hands-on activities using the selected corpus and concordancer. Technical training is necessary to introduce the learners to learning via a computer and more importantly to get them familiar with the interface of the concordancer and the key tools available. Computer literacy is an important requirement for successful corpus consultation as the lack of it and fear of technology can be

detrimental to the success of corpus-based learning, which is entirely computer-based. This training, however, may prove redundant for those who have already mastered the use of computer hardware. Table 8.1 summarises the details of the proposed Technical Training component.

Table 8.1: Summary of the Technical Training Component

Type:	Technical Training
Time:	8 hours 2 hours - Introduction to Computer 6 hours - Corpus Handling
Objectives:	At the end of the training, the students should be able to: 1. use the basic functions of the computer hardware 2. be familiar with the interface of the concordancer 3. differentiate the functions of the key Tools available with the concordancer 4. use the Concordancing Tool to conduct word and phrase-level searches
Method:	Hands-on practices
Teaching Materials	1. Guided questions 2. Practice Worksheets 3. BNC Sampler (written) 4. Corpus Concordancer English (CCE) version 6.5

The next step in the training component is the hands-on practices, whereby learners are guided to conduct simple word-level and phrase-level searches. Concordancer English (CCE) version 6.5 is proposed to be used for this purpose. The objective of this training is to familiarise the learners with the interface of the concordancer by introducing them to its important features and key tools. The CCE has a very simple user interface. It is designed for simple concordancing and collocation searches, rather than complex linguistic investigations and analysis. It has fewer features and tools compared to for instance WordSmith Tools (Scott, 2017) and AntConc (Anthony, 2018). The simple interface has its advantages as it would not be the source of intimidation and confusion to the learners and also makes it easier for the learners to familiarise themselves with the features of the concordancer and their functions. Figure 8.3 below displays is the

interface of CCE taken from LexTutor website at http://www.lextutor.ca/concordancers/concord_e.html:

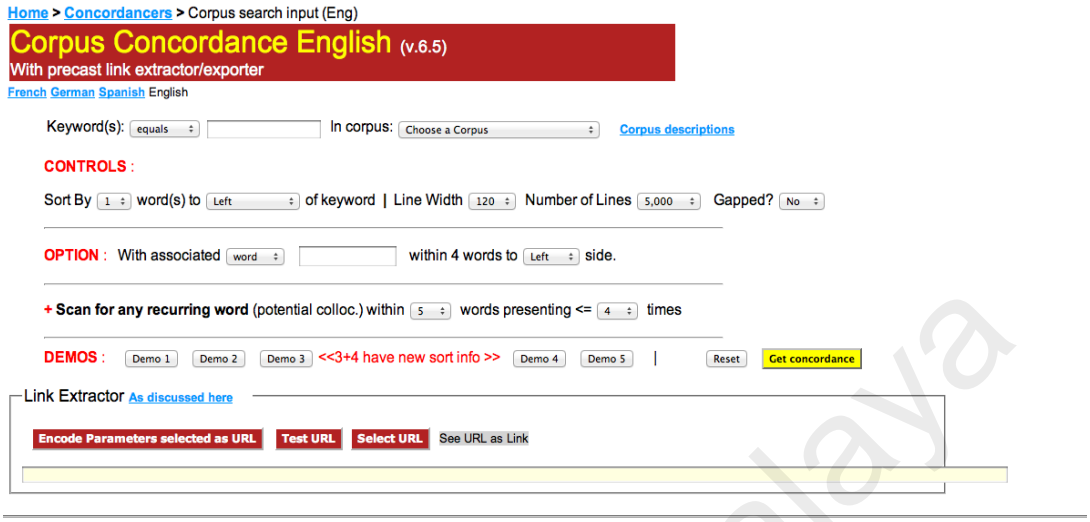


Figure 8.3: Corpus Concordance English Version 6.5

As can be seen in Figure 8.3, CCE is designed only to generate concordances and collocations. The restricted number of functions makes the interface straight forward and user-friendly, which enables learners to use the concordancer with minimum training.

Previous studies indicate that learners find it difficult to make successful searches, and are overwhelmed by the number of results. The proposed training introduces learners to corpus-related information, and trains them to make word and phrase level searches and interpret the results. This involves real searches to get the best results, and to ask the right questions to infer language patterns correctly when interpreting the results. These aspects of the training are essential to prepare the learners for the corpus consultation.

Sun (2000), Cheng et al. (2003), and Vannestal and Lindquist (2007) report that learners have difficulties in processing and interpreting large number of results especially when they were not properly trained in corpus consultation (Sun, 2003; Vannestal & Lindquist, 2007). In view of these difficulties, learners need to be trained in these two aspects of corpus consultation before attempting to self-correct their errors. The

following questions can be used to guide learners to set the search parameters and infer linguistic patterns, and can be modified by training instructors to meet learners' needs and requirements.

Table 8.2: Guiding Questions for Setting the Search Terms and Deducing Language Patterns

Setting search terms	1. What is the form of *BE*? 2. Are you looking for a single verb? or 3. The relation of the verb to other verbs or other parts of speech?
Deducing language patterns	1. Observe the form of the verb after *BE*. 2. Observe the time expressions used. 3. Identify any other auxiliaries used before *BE*. 4. Observe the subjects before *BE*.

It is also suggested that four (4) practise worksheets to be prepared for use in two-hour training sessions. Table 8.3 shows an example of a training worksheet.

Table 8.3: Sample of a Training Worksheet

Language focus: *is +Ving*
Instructions:

1. Search for *is.*
2. Select 10 concordances of the selected verb.
3. Observe the subjects that are used before *is*.
4. Observe the part of speech of the word/phrases after *is + Ving*.
5. Observe any time expression used before and after *is +Ving*.

What kind of subjects can occur before *is +Ving*?	
What part of speech can be used after *is + Ving*?	
When is *is + Ving* used?	

1. In a group of four compare your findings to that of your group members.
2. What conclusions can your group make about the use of *is + Ving*?
3. Finally, share your conclusions with the whole class.

8.2.2.2 Phase 2: Corpus Consultation Component

This section presents the proposal for the main teaching and learning process involved in the corpus consultation component, for which the proposed total time allocated is 10 hours spread over 5 weeks. The proposed corpus consultation is only a part of the ESL writing course, and is not intended to replace the existing practice, but rather to

complement it for a part of the 14 weeks semester. Table 8.4 below provides the summary of the corpus consultation component for the teaching of *BE*.

Table 8.4: Summary of the Corpus Consultation Component

Type:	Corpus consultation
Time:	10 hours
Objectives:	At the end of the teaching and learning sessions, the students should be able to: 1. consult the selected corpus to search for terms needed 2. identify and select relevant examples for the search terms 3. deduce language patterns from the concordance examples 4. propose valid corrections for errors highlighted
Method:	Individual corpus consultation Group work
Teaching Materials/Resources	1. The Four-step Corpus Investigation Guide 2. BNC Sampler (written) 3. Corpus Concordancer English (CCE) version 6.5

Figure 8.4 presents the steps teachers can follow in the integration of corpora in their essay writing classes.

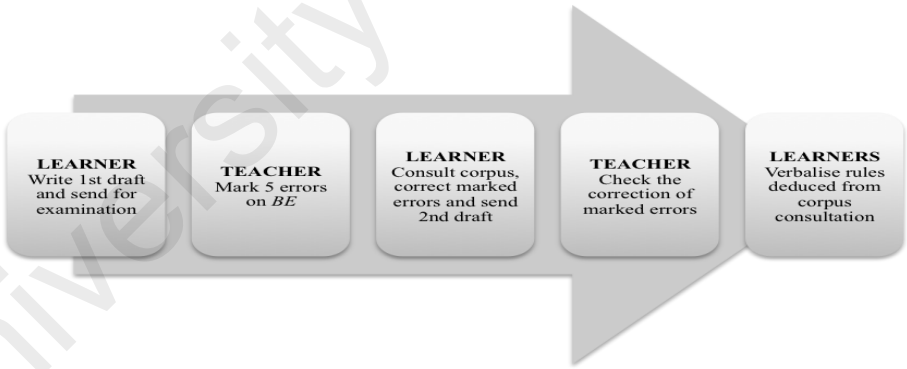


Figure 8.4: Corpus Consultation Process

As illustrated in Figure 8.4 the process begins with the learners write the first draft of their essays as an assignment outside class hours, so that more time can be spent on the actual corpus investigation. When the essays are submitted, the teacher marks them and identifies the five most common errors involving *BE*. Gaskell and Cobb (2004) suggested limiting the number of errors to only five errors per essay as not to burden the learners. The errors are to be highlighted but not coded to maximise the opportunity for

the learners to discover for themselves the types of errors (e.g. tense, agreement, etc.) that are highlighted.

Next, learners are to consult the corpus to correct the marked errors. This session has to be conducted during class time in the presence of the teacher whose task is to assist learners with their searches. Even after the preliminary training sessions, learners are considered beginning corpus users, still needing guidance and facilitation to consult the corpus smoothly and fruitfully. To give learners proper guidance and to avoid losing their way, Kennedy and Miceli (2001) suggest four steps in conducting their searches, namely formulating questions, devising a search strategy, observing and selecting relevant examples, and drawing conclusions. Detailed samples for each step are summarised in Table 8.5.

The next step for the learners is to correct the highlighted errors and write the second draft. In order to keep track of the changes made, learners should be given a form containing information on the patterns and the proposed corrections. The form can be used as an instant means for the teacher to evaluate learners' corrections, keep track of and evaluate their progress, trace their difficulties in corpus investigation and in dealing with the selected grammatical item and most importantly provide statistical evidence for the learners' valid and invalid corrections. In addition, learners would also benefit from the form, as it can be used to evaluate their own progress and get immediate feedback from the teacher on the corrections they have made. It is suggested that the form should be submitted together with the second draft of the essay. Figure 8.5 is a sample of the suggested form.

		Instructor's Remarks
Error 1	:	
Successful search term(s)	:	
Summary of concordance pattern(s)	:	
Proposed correction	:	

Figure 8.5: Error Correction Form

The fourth step requires the teacher to mark the second draft, paying careful attention to the corrections proposed for the five highlighted errors. By using the form accompanying the second draft, the teacher would be able to provide instant feedback, and decide whether learners need to repeat the corpus investigation for inappropriate corrections. If they are successful in the error correction task, they should be ready for the final stage of the corpus consultation process, to explain their findings to their group members.

The final step is added to give learners an opportunity to continuously engage in language-learning process (Vannestal & Linguist, 2007). Stating explicitly and explaining the rule they have found can lead to further linguistic insights (Celce-Murcia & Larsen-Freeman, 2016). Learners' findings from the corpus investigation should be discussed with other group members. This information sharing and awareness raising may help learners understand the patterns they have found. It may also help learners learn something which has been taught, but which they have either forgotten or never really understood in the first place.

Table 8.5: The Four-step Corpus Investigation Guide

Steps	Samples of Actions	Tips
1. Formulate the question	"What is the form of the verb after the highlighted *BE* ?"	• Try to state your question precisely. • Ensure it is specific enough for the situation you are dealing with. • If it is in yes/no or multiple-choice form, consider whether an open question would be more appropriate. For example, rather than asking "Does y come after x?" you might want to ask "What comes after x?" • Keep in mind both lexical and grammatical issues. • In your dealings with *BE*: When considering the correct form of the *BE* construction, look both to the right and to the left, and to a distance of a few words.
2. Devise a search strategy	Search for *BE* Look for examples of *BE* combined with *Ving*. Look for any time expressions used in the *BE + Ving* construction.	• Think about how efficient your strategy will be. Is it likely to generate many irrelevant examples alongside the useful ones? If so, maybe you should restrict your search further. • Check if you are dealing with a variant of a general pattern, with a fixed part and a variable part, as you may want to search only on the fixed part. • If you are not satisfied with the examples found, think about using wildcards or substituting something else for one of the search words: another form of the same lemma or a word that may be equivalent in the context that interests you. • Remember that you can use the dictionary to look for potentially appropriate words.
3. Observe the examples and select relevant ones	Observe the word/phrase following *BE + Ving*. Observe the types of subjects used before *BE*. Observe the use of any other auxiliaries before *BE*. Observe the use of time expressions.	• Remember to check the meaning of examples you want to use as evidence, and seek out those that most closely match the requirements of your target sentence. • Try not to be influenced by assumptions about what you will see in the examples. Look to the left and right of keywords to see which words are linked to them. The words you are expecting to find may not be present, and vice versa. • Try not to be attracted only to the types of usage of a word that occur most frequently. The type you are interested in may be a less common case.
4. Draw conclusions	Identify the combination *BE + Ving* and insert it into the target sentence, making any necessary adaptations.	• Even if you have only one example as evidence, it may be enough on which to base your case. Remember that what matters is how good your evidence is, not how much of it there is. • If you have found only a few examples when you were expecting many, or vice versa, you may need to think about what this means. Why were you expecting to find many or only a few? What has affected the result? • If you have found no examples, think carefully about what conclusion you can draw. Make sure you relate your conclusion to the question that you initially posed.

Adapted from Kennedy and Miceli (2001, p. 82)

8.3 Conclusions

This chapter proposes a corpus consultation model for the teaching and learning of *BE*. The proposal includes suggestions on which area of *BE* are to be highlighted, the choice of corpus and the rationale for the choice, and suitable corpus-based learning requirements of beginners. It also proposes a model for corpus literacy training and the actual corpus consultation sessions. The suggestions presented here are based on successful implementations reported in previous research and the methods and steps have been adopted and adapted to suit ESL learners in Malaysia.

The effectiveness of this model has not been empirically tested and proved in the immediate context. Nevertheless, considering its successful implementation in other ESL settings, the researcher is certain the proposed model, with appropriate implementation can be an effective and interesting method of teaching and learning English, including in particular teaching and learning of *BE*.

As discussed earlier, the model is not designed for semester-long corpus consultation, but extended to complement the existing approach. It is hoped that a short exposure to corpus consultation will help learners extend their knowledge of English, and give them an incentive to continue consulting corpus data outside the language classroom in order to improve their command of other aspects of the language.

More importantly, this chapter addresses the 'so what' question. It provides the application aspect of the study, which is central in any research on learner language. Analysis of what learners can and cannot do with the target language should be followed by the application of the analysis. This thesis follows that thought of logical reasoning. The CCM proposed here is based on the findings of what the L1-Malay learners can and cannot do with *BE*. The findings are the basis for the selection of

aspects of *BE* to include in the teaching component and the corpus-based activity proposed.

REFERENCES

Abdullah, S. & Noor, N.M. (2013). Contrastive analysis of the use of lexical verbs and verb-noun collocations in two learner corpora: WECMEL vs LOCNESS. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world Vol* 1 (pp.139-160). Japan: Kobe University.

Abdul Kader, M.I., Begi, N., Vasegi, R. (2013). A corpus-based study of Malaysian ESL learners' use of modals in argumentative compositions. *English Language Teaching, 6*(9), 146-157.

Abdul Rahim, M. E., Abdul Rahim, E. M. & Chia, H.N. (2013). Distribution of articles in written composition among Malaysian ESL learners. *English Language Teaching, 6*(10), 149-157.

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 35-53). Amsterdam & Atlanta: Rodopi.

Aijmer, K. (2002). Modality in advance Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and language teaching* (pp. 57-76). Amsterdam: John Benjamins.

Akande, A. T. (2013). Non-standard syntactic features in Nigerian university graduates' English. *Awka Journal of English Language and Literary Studies, 4*(1), 16-29.

Ali, A. M. (2007). Semantic fields of problem in business English: Malaysian and British journalistic business texts. *Corpora, 2*(2), 211-239.

Allwright, D., Bailey, K.M., (1991). *Focus on the language classroom: An introduction to classroom research for language teachers.* Cambridge: Cambrige University Press.

Altenberg, B. & Granger, S. (2001). The grammatical and lexical patterning of *MAKE* in native and non-native student writing. *Applied Linguistics, 22*(2), 173-195.

Ambridge, B., Rowland, C., Theakston, A., & Tomasello, M. (2006). Comparing different accounts of inversion errors in children's non-subject wh-questions: 'What experimental data can tell us?'. *Journal of Child Language, 33*(3), 519-551.

Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Arjan, A., Abdullah, N.H. & Roslin, N. (2013). A corpus-based study on English prepositions of place, *in* and *on*. *English Language Teaching, 6*(12), 167-174.

Arshad, S. (2002). The English of Malaysian School Students (EMAS) Corpus. Available at: http://works.bepress.com/arshad_abdsamad/2/

Arshad, S. (2004). Beyond concordance lines: Using concordances to investigating language development. *Internet Journal of e-Language Learning & Teaching, 9*(1), 43-51.

Arshad, S., & Hawanum, H. (2010). Teaching grammar and what student errors in the use of the English auxiliary 'be' can tell us. *The English Language Teacher, 39*, 164-178.

Aziz, R. A. & Mohd Don, Z. (2013). The *BE* verb omissions among advanced L1-Malay ESL learners: What corpus-based study can reveal. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world Vol 1* (pp.121-138). Japan: Kobe University.

Aziz, R. A. & Mohd Don, Z. (2014). The overgeneration of *be+verb* in the writing of L1-Malay ESL learners in Malaysia. *Research in Corpus Linguistics 2*, 35-44. http://www.aelinco.es/ojs/index.php/ricl/article/view/28

Aziz, R. A., Jin, C. C. & Nordin, N. M. (2016). The use interactional metadiscourse in the construction of gender identities among Malaysian ESL learners. *3L: The Southeast Asian Journal of English Language Studies, 22(*1), 207-220.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society, 19*(3), 273-306.

Baker, P., Gabrielatos, C., & McEnery, T. (2013a). *Discourse analysis and media attitudes: The representations around the word Islam in the British press*. Cambridge: Cambridge University Press.

Baker, P., Gabrielatos, C., & McEnery, T. (2013b). Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim'in the British press 1998–2009. *Applied Linguistics, 34*(3), 255-278.

Balcom, P. (1997). Why is this happened? Passive morphology an unaccusativity. *Second Language Research, 13*(1), 1-9.

Barabadi, E., & Khajavi, Y. (2017). The effect of data-driven approach to teaching vocabulary on Iranian students' learning of English vocabulary. *Cogent Education, 4*(1), 1283876.

Barlow, M. (2005). Computer-based analyses of learner language. In R. Ellis & G. Barkhuizen (Eds.), *Analysing Learner Language* (pp. 335-357). Oxford: Oxford University Press.

Becker, M. (2004). Copula omission is a grammatical reflex. *Language Acquisition, 12*(2), 157-167.

Becker, M. (2002). The development of copula in child English. The lightness of *be*. *Annual Review of Language Acquisition, 2*, 37-58.

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp.15–36). Amsterdam, Netherlands: John Benjamins.

Biber, D. (1986a). On the investigation of spoken/written differences. *Studia Linguistica, 40,* 1-21.

Biber, D. (1986b). Spoken and written textual dimension in English: Resolving the contradictory findings. *Language Sciences, 62,* 384-414.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2003). Compressed noun-phrase structures in newspaper discourse. In J. Aitchison & D. M. Lewis (Eds.), *New media discourse* (pp.169-181). London: Routledge.

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics, 14*(3), 275-311.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English.* Harlow, Essex: Longman.

Biber, D., Conrad, S. M., & Reppen, R. (1994). Corpus-based approach to issues in applied linguistics. *Applied Linguistics, 15*(2), 169-190.

Biber, D., Conrad, S. M., & Reppen, R. (1998). *Corpus linguistics. Investigating language structure and use.* Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P. Helt, M. (2002). Speaking and writing in the universities. A multidimensional comparison. *TESOL Quarterly, 36*(1), 9-48.

Biber, D., & Finegan, E. (1988). Drift in three English genres. In T. J. Walsh (Ed.), *Synchronic and diachronic approaches to linguistic variation and change* (pp. 22-36). Washington DC: Georgetown University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Pearson Education Limited.

Bloom, L. (1970). *Language development: Forms and functions of emerging grammars*. Cambridge, MA: MIT Press.

Botley, S. P. (2010). A corpus-based comparison of idiom use by Malaysian, British and American students. In *Proceeding of International Conference on Science and Social Research* (pp.139-144).

Botley, S. P. (2014). Argument structure in learner writing: A corpus-based analysis using argument mapping. *Kajian Malaysia, 32* (1), 45-77.

Botley, S. P. & Dillah, D. (2007). Investigating spelling errors in a Malaysian learner corpus. *Malaysian Journal of ELT Research, 3,* 74-93.

Botley, S. P., De Alwis, C., Metom, L., & Izza, I. (2005). *CALES: A corpus-based archive Of learner English in Sarawak. Final project report*: Unit for Research, Development and Commercialisation, Universiti Teknologi MARA.

Botley, S. P., Haykal, H. Z. & Monalisa, S. (2005). Lexical and grammatical transfer by Malaysian student writers. In *the Proceeding of 10th International Conference on Translation, Universiti Malaysia Sabah, Kota Kinabalu, 2nd – 4th August.*

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning, 60*(3), 534–572.

Boulton, A. (2011). Data-driven learning: The perpetual enigma. In S. Gazdz-Roszkowski (Ed.), *Explorations across languages and corpora* (pp. 563-580). Frankfurt: Peter Lang.

Brown, R. (1973). *A first language: The early stages*. Cambridge, Massachusetts: Harvard University Press.

Brindle, A. (2015). A corpus analysis of discursive constructions of the sunflower student movement in the English-language Taiwanese press. *Discourse & Society, 27*(1), 3-19.

Burzio, L. (1986). *Italian syntax: A government-binding approach*. Dordrecht: In Reidel.

Callies, M. (2009). 'What is even more alarming is...' - A contrastive learner-corpus study of what-clefts in advanced German and Polish L2 writing. In M. Wysocka (Ed.), *On language structure, acquisition and teaching. Studies in honour of Janusz Arabski on the occasion of his 70th birthday* (pp. 283-292). Katowice: Wydawnictwo Uniwersyetu Slaskiego.

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics, 16*, 141-158.

Celce-Murcia, M., & Larsen-Freeman, D. (2016). *The grammar book: An ESL/EFL teacher's course* (2nd ed.). USA: Heinle & Heinle Publishers.

Chan, Y. W. (2004). Syntactic transfer: Evidence from the interlanguage of Hong Kong Chinese ESL learners. *The Modern Language Journal, 88*, 56-74.

Chan, T., Albakry, M., Williams, R., Lamb, B., Kelsey, D., van Dijk, T., ... & Owens, J. (2017). The umbrella movement in the media: A corpus-driven analysis of newspapers in Hong Kong and China. *Journalism and Discourse Studies*, 2-2.

Channell, J. (2000). Corpus-based analysis of evaluative texts. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 38-55). Oxford: Oxford University Press.

Chambers, A. and O'Sullivan, Í. (2004) Corpus consultation and advanced learners' writing skills in French. *ReCALL, 16*(1), 158–172.

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes, 6*(4), 289–302.

Cheng, W., Warren, M., & Xun-Feng, X. (2003). The language learner as language researcher: Putting corpus linguistics on the timetable. *System, 31*(2),173–186.

Cheng, W., & Lam, P. W. (2013). Western perceptions of Hong Kong ten years on: A corpus-driven critical discourse study. *Applied Linguistics, 34*(2), 173-190.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, N. (1962). In *the 3rd Texas Conference Problems in Linguistic Analysis in English University of Texas, Austin.*

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology, 11*, 38-63.

Collins, P. (1994). Extraposition in English. *Functions of Language, 1*, 7-24.

Connor, U. (2004). Intercultural rhetoric research: Beyond texts. *Journal of English for Academic Purposes, 3*, 291–304.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics, 22*(-1), 75-95.

Conrad, S. M. (2001). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In S. M. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies*. Harlow: Longman.

Conrad, S. M. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century?. *TESOL Quarterly, 34*, 548-559.

Conrad, S. M. (1996). Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education, 8*(3), 299-326.

Dana, W. (2008). Differences in men's and women's ESL academic writing at the University of Melbourne. *Jurnal Sosioteknologi, 14* (7), 447-463.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System, 26*, 163-174.

de Beaugrande, R. (1996). The 'pragmatics' of doing language science: The 'warrant' for large-corpus linguistics. *Journal of Pragmatics, 25*(4), 503-535.

de Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, New Series 2, 225-246.

de Haan, P. (1992). The optimum corpus sample size? In G. Leitner (Ed.), *New directions in English language corpora* (pp. 3-19). Berlin and New York: Mouton de Gruyter.

de Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE Tagger. In C. Mair & M. Hundt (Eds.), *Corpus Linguistics and linguistic theory* (pp. 69-79). Amsterdam: Rodopi.

Díaz-Negrillo, A., Meurers, D., Valera, S. & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, *36*(1-2), 139-154.

Dulay, H. & Burt, M. (1974). A new perspective on the creative construction processes in child second language acquisition. In Ellis, R. (2008). *The study of second language acquisition 2nd edition* (p. 53). Oxford: Oxford University Press.

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two*. Oxford: Oxford University Press.

Eaton, H. (1940). *Semantic frequency list of English, French, German and Spanish*. Chicago: Chicago University Press.

Ellis, R. (1985). *Understanding second language acquisition.* Oxford: Oxford University Press.

Ellis, R. (1988). The effects of linguistic environment on the second-language acquisition of grammatical rules. *Applied Linguistics*, *9*(3), 257-74.

Ellis, R. (2008). *The study of second language acquisition 2nd edition*. Oxford: Oxford University Press.

Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness, 17*(1), 25-43.

Fleta, T. M. (2003). Is-insertion in L2 Grammars of English: A step forward between developmental stages*?* In *the Proceeding of the 6th Generative Approaches to Second Language Acquisition Conference (GASLA 2002)* (pp. 87-96).

Flowerdew, L. (1998). Corpus linguistic techniques applied to textlinguistics. *System, 26*(4), 541-552.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes, 24*(3), 321-332.

Franchis, G. (1994). Labelling discourse: An aspect of nominal-group cohesion. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 83-101). New York: Routledge.

Fries, C. (1952). *The structure of English: An introduction to the construction of sentences*. New York: Harcourt-Brace.

Friginal, E., Man Li & Weigle, S.C. (2014). Revisiting multiple profiles of learners compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing, 23*, 1-16.

Garside, R., Leech, G., & McEnery, T. (1997). *Corpus annotation: Linguistic information from computer text corpora*. London and New York: Longman.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp.137-150). London: Longman.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors?. *System, 32,* 301–319.

Gavruseva, E., & Meisterheim, M. (2003). On the syntax of predication in child L2 English. In *the Proceeding of the 6th Generative Aprroaches to Second Language Acquisition Conference* (pp.115-121).

Gilquin, G. (2012). Lexical infelicity in English causative constructions. Comparing native and learner collostructions. In J. Leino & R. von Waldenfels (Eds.), *Analytical causatives. From 'give' and 'come' to 'let' and 'make'* (pp. 41-63). Munchen: Lincom

Gilquin, G. & Granger, S. (2015). Learner language. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 418-435). Cambridge University Press: Cambridge.

Gilquin, G., de Cock, S. & Granger, S. (2010). Louvain international database of spoken English interlanguage. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Granath, S. (2009). Who benefits from learning how to use corpora? In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 47–65). Amsterdam: John Benjamins.

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 37-51). Amsterdam: Rodopi.

Granger, S. (1998a). The computer learner corpus. The versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). New York: Longman.

Granger, S. (1998b). *Learner English on computer*. London and New York: Addison Wesley Longman.

Granger, S. (2002). A Bird's-eye view of learner corpus reseach. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam: John Benjamins

Granger, S. (2003). Error-tagged learner corpora and CALL: A promisiong synergy. *CALICO Journal*, *20*(3), 465-480.

Granger, S., Dagneaux, E., & Meunier, F. (2002). *The International Corpus Of Learner English Handbook And CD-ROM*. Louvain-la Neuve: Presses Universitaires de Louvain.

Granger, S. & Rayson, P. (1998). Automatic lexical profiling of learner texts. In S. Granger (Ed.), *Learner English on Computer* (pp. 119-131). London & New York: Addison Wesley Longman.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL. *World Englishes*, *15*, 19-29.

Granger, S., Paquot, M., & Rayson, P. (2006). Extraction of multiword units from EFL and native English corpora. The phraseology of the verb 'make'. In A. Hacki Buhofer and H. Burger (Eds.), *Phraseology in motion I: Methoden und kritik*. (pp. 57-68). Baltmannsweiler: Schneider Verlag Hohengehren

Guan, X. (2013). A study on the application of data-driven learning in vocabulary teaching and leaning in China's EFL Class. *Journal of Language Teaching & Research, 4*(1), 105-112.

Halliday, M. A. K. (1993c). Quantitative studies and probabilities in grammar. In M. Hoey (Ed.), *Data, description, discourse. Papers on the English language in honour of John McH. Sinclair* (pp.1-25). London: HarperCollins.

Hasselgård, H. (2009). Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer (Ed.) *Corpora and language teaching* (pp. 120-39). Amsterdam & Philadelphia: John Benjamins.

Hawkins, R., & Casillas, G. (2008). Explaining frequency of verb morphology in early L2 speech. *Lingua, 118*(4), 595-612.

Haznedar, B. (2001). The acquisition of the IP system in child L2 English. *SSLA, 23*, 1-39.

Haznedar, B. (2007). The acquisition of tense-aspect in child second language English. *Second Language Research, 23*(4), 383-417.

Haznedar, B., & Schwartz, B. (1997). Are there optional infinitives in child L2 acquisition?. In *the 21st Annual Boston University Conference on Language Development, Somerville, MA.*

Herat, M. (2005). 'BE' variation in Sri Lankan English. *Language Variation and Change, 17,* 181-208.

Herriman, J. (2000a). Extraposition in English: A study of the interaction between the matrix predicate and the type of extraposed clause. *English Studies, 81,* 582-599.

Hinkel, E. (2004). Tense, aspect and passive voice in L1 and L2 academic texts. *Language Teaching Research, 8*(1), 5-29.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quaterly, 37*(2), 275-301.

Hinkel, E. (2002). *Second language writers' text. Linguistic and rhetorical features*. London: Lawrence Erlbaum Associates.

Hirakawa, M. (2006). Passive unaccusative errors in L2 English revisited. In Slabakova, R., Montrul, S., Prevost, P. (Eds.), *Inquiries in linguistic development: In honour of Lydia White* (pp. 17-39). Amsterdam: John Benjamins.

Ho, M. L., & Platt, J. T. (1993). *Dynamics of a contact continuum: Singaporean English*. Oxford: Clarendon.

Holm, J. (1984). Variability of the copula in Black English and its creol kin. *American Speech, 59*(4), 291-309.

Hong, A.L, Rahim, H.A., Hua, T.K. & Salehuddin, K. (2011). Collocations in Malaysian English learners' writing: A corpus-based error analysis. *The Southern Asian Journal of English Language Studies, 17*, 31-44.

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 77-116). Amsterdam: John Benjamins.

Hughes, R., & McCarthy, M. (1998). From sentence to discourse: Discourse grammar and English language teaching. *TESOL Quaterly, 32*, 263-287.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2002). Pattern grammar, language teaching, and linguistic variation. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 167-183). Amsterdam: John Benjamins.

Hunston, S., & Franchis, G. (2000). *Pattern grammar*. Amsterdam: John Benjamins.

Hyland, K., (2005). *Metadiscourse*. Continuum: London.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics, 25*(2), 156-177.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*(2), 183-205.

Ionin, T., & Wexler, K. (2001). L1 Russian children learning English: Tense and overgeneration of 'be'. In *the Proceeding of Second Language Research Forum 2000*.

Ionin, T., & Wexler, K. (2002). Why is 'is' easier than '-s'?: Acquisition of tense/agreement morphology by child second language learners of English. *Second Language Research, 18*(2), 95-136.

Izumi, E., Uchimoto, K., & Isahara, H. (2005). Error annotation for corpus of Japanese Learner English. In *the 6th International Workshop on Linguistically Annotated Corpora, Jeju Island, Korea.*

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377–403.

Jishvithaa, J. M., Tabitha, M., & Kalajahi, S. A. R. (2013). Teaching grammar: The use of the English auxiliary 'BE' present tense verb among Malaysian form 4 and form 5 students. *Advance in Language and Literary Studies, 4*(2), 152-158.

Joharry, S. A. (2013). Corpus-based study on Malaysian users' English writing: A preliminary study of the collocational behaviour and semantic prosody of cause. In *Third Malaysian Postgraduate Conference (MPC) 2013* (pp. 89-96).

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *ELR Journal, 4*, 1-16.

Johns, A. (1997). *Text, role, and context: Developing academic literacies*. Cambridge: Cambridge University Press.

Johns, T. (1997) Contexts: the background, development and trialling of a concordance-based CALL program. In Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (Eds.), *op. cit*., 100-115.

Ju, M. K. (2000). Overpassivization errors by second language learners. *Studies in Second Language Acquisition, 22*, 85-111.

Kafipour, R., & Khojasteh, L. (2012). A comparative taxonomy of errors made by Iranian undergraduates learners of English. *Canadian Social Science, 8*(1), 18-24.

Kamarudin, R. (2013). A corpus-based study on the use of phrasal verbs by Malaysian learners of English: The case of particle *UP*. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world Vol 1* (pp.255-270). Japan: Kobe University.

Kanestion, A., Singh, M. K. S., Shamsudin, S., Isam, H., Kaur, N., & Singh, G. S. P. (2016). Lexical verbs in Malaysian University English Test argumentative essays: A corpus-based structural analysis. *International Review of Management and Marketing, 6*(8S), 13-17.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology, 5*(3), 77-90.

Kettemann, B. (1995). On the use of concordancing in ELT. *TELL & CALL , 4*, 4-15.

Knowles, G., & Zuraidah, M. D. (2004). Introducing MACLE: The Malaysian Corpus Of Learner English. In *the 1st National Symposium of Corpus Linguistics and Foreign Language Education*.

Knowles, G., Zuraidah, M. D., Jariah, M. J., Rajeswary, S., Janet, Y., Sathiadevi, et al. (2006). The Malaysian Corpus of Learner English: A bridge from linguistics to ELT. In H. Azirah & H. Norizah (Eds.), *Varieties of English in Southeast Asia and beyond*. Kuala Lumpur: University of Malaya Press.

Kotamjani, S. S., Razavi, O. F., & Hussin, H. (2017). Online corpus tools in scholarly writing: A case of EFL postgraduate student. *English Language Teaching, 10*(9), 61-68.

Lakshmanan, U. (1995). Child second language acquisition of syntax. *Studies in Second Language Acquisition, 17*, 201-229.

Lardiere, D. (1998). Dissociating syntax from morhology in divergent L2 end-state grammar. *Second Language Research, 14*(4), 359-375.

Lee, N., & Huang, Y. Y. (2004). To be or not to be- The variable use of the verb '*be'* in the interlanguage of Hong Kong Chinese children. *Regional Language Centre Journal, 32*(2), 211-228.

Lee, D. & Swales, J. A. (2006). Corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*(1), 56-75.

Leel, H. (2011). In defense of concordancing: An application of data-driven learning in Taiwan. *Procedia Social and Behavioral Sciences,12*, 399–408.

Leech, G. (1999). The distribution and function of vocatives in American and British English conversation. *Language and Computers, 26*, 107-120.

Leech, G. (1998). Learner corpora: What they are and what can be done with them. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xix). London and New York: Addison Wesley Longman.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 105-122). Berlin: Mouton de Gruyter.

Leki, I. (1999). *Academic writing. Techniques and tasks* (3rd ed.). New York: Cambridge University Press.

Levin, B., & Rappaport Hovav, M. (1995). *Unaccusativity: At the syntax-lexical semantics interface*. Cambridge, Massachusetts: MIT Press.

Loke, D. L., Ali, J., & Zulkifli Anthony, N. N. (2013). A corpus based study on the use of preposition of time 'on' and 'at' in argumentative essays of form 4 and form 5 Malaysian students. *English Language Teaching, 6*(9), 128-135.

Longman dictionary of contemporary English. (2015). Essex: Pearson Education Limited.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosidies. In M. Baker, G. Franchis & E. Tognini-Bonelli (Eds.), *Text and technology* (pp. 157-176). Amsterdam: Benjamin.

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk, 3rd Edition.* Mahwah, NJ: Erlbaum.

Makino, T., (1993). Learner self-correction in ESL written compositions. *ELT Journal, 47*, 337-341.

Manokaran, J., Ramalingam, C., & Adriana, K. (2013). A corpus-based study on the use of past tense auxiliary 'BE' in argumentative essays of Malaysian ESL learners. *English Language Teaching, 6*(10), 111-119.

Marín Arrese, J. I. (2015). Epistemicity and stance: A cross-linguistic study of epistemic stance strategies in journalistic discourse in English and Spanish. *Discourse Studies*, *17*(2), 210-225.

Maros, M., Tan, K. H., & Khazriyati, S. (2007). Interference in learning English: Grammatical errors in English writing among rural Malay secondary school students in Malaysia. *Jurnal e-Bangi, 2*(2), 1-15.

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

McCrostie, J. (2008). Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME Journal, 32*, 97-114.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics*. United Kingdom: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics. 2nd Edition*. Edinburgh: Edinburg University Press.

McEnery, T., Xiao, T., & Tono, Y. (2006). *Corpus-based language studies*. London: Routledge.

Meunier, F. and de Mönnink, I. (2001). Assessing the success rate of EFL learner corpus tagging. In *ICAME Conference, Louvain-la-Neuve, Spain*.

Miceli, T., & Kennedy, C. (2002). An Apprenticeship with the CWIC Corpus: A tool for learner writers in Italian. In *Proceedings of Workshop Innovations in Italian Teaching Brisbane, Griffith University* (pp. 83-94).

Milton, J. (2001). Elements of a written interlanguage: A computational and corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese Students. In G. James (Ed.), *Research report Vol 2*. Language Centre: The Hong Kong University of Science and Technology.

Mohd Don, Z. & Sriniwass, S. (2017). Conjunctive adjuncts in Malaysian undergraduate ESL Essays: Frequency and manner of use. *Moderna Sprak, 1*, 99-117.

Moscati, V. (2006). Parameterizing negation: Interactions with copula constructions in Italian and English children. In B. Belleti & C. D. D. Ferrari (Eds.), *Language acquisition and development* (pp. 367-378). Cambridge: Cambridge Scholar Press.

Mukundan, J., Saadullah, K. A., Ismail, R., & Jusoh Zasenawi, N. H. (2013). Malaysian ESL students' syntactic accuracy in the usage of English modal verbs in argumentative writing. *English Language Teaching, 6*(12), 98-105.

Mukundan, J., & Rezvani Kalajahi, S. A. (2013). *Malaysian Corpus of Student Argumentative Writing*. Australia: Australian International Academic Centre.

Muneera Yahya, A. M., & Wong, B. E. (2011). The acquisition of English *'be'* auxiliary and thematic verb constructions by adult Arab ESL learners. *International Journal of English Linguistics, 1*(2), 91-102.

Murad, T. M., & Khalil, M. H. (2015). Analysis of errors in English writings committed by Arab first-year college students of EFL in Israel. *Journal of Language Teaching and Research, 6*(3), 475 - 481.

Muthusamy, P. & Farashaiyan, A. (2017). A corpus-based comparative study on Malaysian ESL learners and native English speakers in compliment patterns. *International Journal of Linguistics, 9*(5), 232-246.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research, 21*(4), 373-391.

Nagata, R., Whittaker, E., & Sheinman, V. (2011). Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 1210-1219).

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125-152). Amsterdam: John Benjamins.

Nik Safiah, K., Farid, M. O., Hashim, H. M., & Abdul Hamid, M. (2010). *Tatabahasa dewan. Edisi ketiga*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Nor Hashimah, J., Norsimah, M. A., & Kesumawati, A. B. (2008). The mastery of English language among lower secondary school students in Malaysia: A linguistic analysis. *European Journal of Social Sciences, 7*(2), 106-119.

Odlin, T. (2003). Cross-linguistic influence. In C.H. Doughty & M.H. Long (Eds.), *The handbook of L2 acquisition* (pp. 436-486). London: Blackwell.

O'Keeffe, A., McCarthy, M. J. & Carter, R. A. (2007). *From corpus to classroom*. Cambridge: Cambridge University Press.

Olofsson, A. (2004). Them bones, them bones... Why is the leg bone connected *to* rather than *with* the knee bone? *Gothenburg Studies in English, 88*, 163-180.

Oshita, H. (2000). What is happened may not be what appears to be happening: A corpus study of 'passive' unaccusatives in L2 English. *Second Language Research, 16*(4), 293-324.

O'Sullivan, I. (2007). Enhancing a process oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL, 19*(3), 269-286.

O'Sullivan, Í. and Chambers, A. (2006) Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing, 15*(1), 49–68.

Park, K. S., & Lakshmanan, U. (2007). The unaccusative-unergative distinction in resultatives: Evidence from Korean L2 learners of English. In *Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition North America (GALANA)* (pp. 328-338).

Partington, A. (1998). *Patterns and meanings*. Amsterdam/Phidelphia: John Benjamins.

Pérez-Paredes, P., Jiménez, P. A., & Hernández, P. S. (2017). Constructing immigrants in UK legislation and Administration informative texts: A corpus-driven study (2007–2011). *Discourse & Society*, *28*(1), 81-103.

Perlmutter, D. M. (1978). Impersonal passives and the unaccusative hypothesis. In *Annual Meeting of the Berkeley Linguistics Society* (pp. 157-190).

Petch-Tyson, S. (2000). Demonstrative expressions in argumentative discourse. A computer based comparison of non-native and native English. In S. Botley & A. M. McEnery (Eds.), *Corpus-based and computational approaches to discourse anaphora* (pp. 43-64). Amsterdam & Philadelphia: John Benjamins.

Phoocharoensil, S. (2012). Language corpora for EFL teachers: An exploration of English grammar through concordance lines. *Procedia - Social and Behavioral Sciences, 64*, 507-514.

Pine, J., Conti-Ramsden, G., Joseph, K., Lieven, E., & Serratrice, L. (2008). Tense over time: Testing the Agreement/Tense Omission Model as an account of the pattern of tense-marking provision in early child English. *Journal of Child Language, 35*(1), 55-57.

Platt, J., & Weber, H. (1980). *English in Singapore and Malaysia. Status, features, functions.* Kuala Lumpur: Oxford University Press.

Prevost, P., & White, L. (1999). Finiteness and variability in SLA: More evidence for missing surface inflection. In *the 23rd Annual Boston University Conference on Language Development*.

Prevost, P., & White, L. (2000). Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research, 16*(2), 103-133.

Preyer, W. (1889). *The mind of a child*. New York: Appleton.

Rayson, P., & Wilson, A. (1996). The ACAMRIT semantic tagging system: Progress report. In *Proceedings of Language Engineering for Document Analysis and Recognition, LEDAR, AISB96 Workshop* (pp. 13-20).

Rice, M. L., Wexler, K., & Hershberger, S. (1998). Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech, Language and Hearing Research, 43*, 1126-1145.

Richards, J. (1971). A non-contrastive approach to error analysis. *ELT Journal, 25*, 204-19.

Romer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics, 31*, 205–225.

Sabet, M. K. & Minaei, R. (2017). A comparative corpus-based analysis on specific discourse: The quantitative and qualitative academic papers in the field of the TEFL. *Theory and Practice in Language Studies, 7*(4), 294-304.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 129-158.

Schütze, C. (2004). The non-omission of nonfinite "be". *Nordlyd*, *31*(3).

Scollon, R. (1994). As a matter of fact: The changing ideology of authorship and responsibility in discourse. *World Englishes, 13*, 33-46.

Scott, M. (1998). *WordSmith tools manual*. University Press.

Scott, M. (2008). WordSmith Tools (Version 5). Liverpool: Lexical Analysis Software.

Scott, M. (2017). WordSmith Tools (Version 7). Stroud: Lexical Analysis Software.

Scovel, T. (2001). *Learning new languages: A guide to second language acquisition.* Massachsetts: Heinle & Heinle.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10*, 209-231.

Shi, J. (2017). Biting off more than they can chew? The impact of pedagogical application of corpus on vocabulary ability of intermediate-level ESL learners in mainland China: A quasi-experimental study. *English Language Teaching, 10*(9), 232-244. doi: 10.5539/elt.v10n9p232

Sinclair, J. (1991). *Corpus, concodance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1996). EAGLES. Preliminary recommendations on Corpus Typology. http://www.ilc.pi.it/EAGLES96/corpustyp/corputyp.html

Siti Hamin, S., & Mohd Mustafa, I. (2010). Analysis of errors in subject-verb agreement among Malaysian ESL learners. *The Southeast Asian Journal of English Language Studies*, *16*(1), 1-18.

Staples, S. & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing*, *32*, 17-35.

Stern, W. (1924). *Psychology of early childhood up to six years of age*. New York: Holt.

Stevens, V. (1995) Concordancing with language learners: Why? When? What? *CÆLL Journal*, *6*(2), 2-10. http://www.eisu.bham.ac.uk/johnstf/stevens.htm.

Stubbs, M. (1993). *British traditions in text analysis. From Firth to Sinclair*. Amsterdam/Philadelphia: John Benjamins.

Sun, Y. C. (2000). Using on-line corpus to facilitate language learning. In *The Annual Meeting of the Teachers of English to Speakers of Other Languages.*

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M., & Feak, C. B. (2012). *Commentary for Academic writing for graduate students: Essential tasks and skills*. Ann Arbor: Michigan University Press.

Tadros, A. (1994). Predicative categories in expository text. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 69-82). New York: Routledge.

Theakston, A., & Lieven, E. (2008). The influence of discourse context on children's provision of auxiliary 'BE'. *Journal of Child Language, 35*(1), 129-158.

Theakston, A. L., Lieven, E., Pine, J., & Rowland, C. F. (2000). The role of performance limitations in the acquisition of 'mixed' verb arqument structure at Stage 1. In M. Perkins & S. Howard (Eds.), *New directions in language development and discorders* (pp. 119-128). New York: Kluwer Academic/Plenum.

Theakston, A. L., & Rowland, C. F. (2009). The Acquisition of auxiliary syntax: A longitudinal elicitation study. Part 1: Auxiliary 'BE'. *Journal of Speech, Language & Hearing Research, 52*(6), 1449-1470.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal, 97*(1), 77-101

Thurstun, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes, 17*(3), 267–280.

Ting, S. H., Mahanita, M., & Chang, S. L. (2010). Grammatical errors in spoken English of university students in oral communication course. *GEMA Online™ Journal of Language Studies, 10*(1), 53-70.

Todd, W. R. (2001). Induction from self-selected concordances and self-correction. *System, 29*, 91-102

Tode, T. (2003). From unanalyzed chunks to rules: The learning of the English copula '*be*' by beginning Japenese learners of English. *International Review of Applied Linguistics in Language Teaching, 41*(1), 23-53.

Tode, T. (2007). Durability problems with explicit instruction in an EFL context: The learning of the English copula *be* before and after the introduction of the auxiliary *be*. *Language Teaching Research, 11*(1), 11-30.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work* (Vol. 6). Amsterdam/Philadelphia: John Benjamins.

Tono, Y. (2003). Learner corpora: Design, development and aplications. In *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 800-809).

Torgersen, E. N., Gabrielatos, C., & Hoffmann, S. (2011). A corpus-based study on pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory, 7*(1), 93-118.

Tottie, G., & Hoffmann, S. (2006). Tag questions in British and American English. *Journal of English Linguistics, 34*(4), 283-311.

Turnbull, J., & Burston, J. (1998). Towards independent concordance work for students: Lessons from a case study. *ON-CALL, 12*(2), 10-21.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of English language*. London: Longman.

Unlu, E. A., & Hatipoglu, C. (2012). The acquisition of the the copula *be* in present simple tense in English by native speakers of Russian. *System, 40*, 255-269.

Upton, T. A., & Connor, U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes, 20*(4), 313-329.

van Lier, L. (1996). *Interaction in the language curriculum: Awareness, autonomy and authenticity.* London: Longman.

van Rooy, B. & Schäfer, L. (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies, 20*, 325-335. doi:10.2989/16073610209486319

Vannestål, M. E. & Lindquist, H. (2007). Learning English grammar with a corpus: Experimenting with concordancing in a university grammar course. *ReCALL, 19*(3), 329-350.

Vedler, Z. (1967). *Linguistics in philosophy*. Ithaca: Cornell University Press.

Vyatkina, N. (2017). Data-driven learning of collocations: Learner performance, proficiency, and perceptions. *Language Learning & Technology, 20*(3), 159-179.

Wee, R. (2009). Sources of errors: An interplay of interlingual influence and intralingual factors. *European Journal of Social Sciences, 11*(2), 349-359.

Wee, R., Sim, J., & Kamaruzaman, J. (2010). Verb-from errors in EAP writing. *Educational Research and Review, 5*(1), 16-23.

Wexler, K., Schütze, C. T., & Rice, M. (1998). Subject case in children with SLI and unaffected controls: Evidence for the Agr/Tns omission model. *Language Acquisition, 7*(2-4), 317-344.

White, L. (1989). *Universal grammar and second language acquisition*. Amsterdam: John Benjamins.

Wichmann, A. (2004). The intonation of please-requests: A corpus-based study. *Journal of Pragmatics, 36*(9), 1521-1549.

Wilson, S. (2003). Lexically specific constructions in the acquisition of inflection in English. *J. Child Language, 30*, 75-115.

Yip, V. (1994). *Interlanguage and learnerbility: From Chinese to English*. Amsterdam: John Benjamins.

Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology, 12*(2), 31–48.

Yoon, H. (2011). Concordancing in L2 writing class. An overview of research and issues. *Journal of English for Academic Purposes, 10*, 130-139.

Yoon, H. & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing, 13*, 257-283.

Yunus, K. & Awab, S. A. (2011). Collocational competence among Malaysian undergraduate law students. *Malaysian Journal of ELT Research, 7*(1), 151-202.

Yunus, K. & Awab, S. A. (2012). The effects of the use of module-based concordance materials and data-driven learning (DDL) approach in enhancing the knowledge of collocations of prepositions among Malaysian undergraduate Law students. *International Journal of Learning, 18*(9), 181-197.

Yunus, K. & Awab, S. (2014). The impact of data-driven learning instruction on Malaysian law undergraduates' colligational competence. *Kajian Malaysia*, *32*(1), 79-109.

Zhang, G. (2015). It is suggested that… or it is better to…? Forms and meanings of subject it-extraposition in academic and popular writing. *Journal of English for Academic Purposes, 20*, 1-13.

Zarifi, A., & Mukundan, J. (2014). Creativity and unnaturalness in the use of phrasal verbs in ESL learner language. *The Southeast Asian Journal of English Language Studies, 20*(3), 51-62.

Zobl, H. (1989). Canonical typological structures and ergativity in English L2 acquisition. In S. M. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 203-221). Cambridge: Cambridge University Press.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

A. List of Publications

No	Title of Publication	Type of Publication	Year
1.	The ***BE*** verb omission among advanced L1-Malay ESL learners: What Corpus-based study can reveal. *Roslina Abdul Aziz* *Zuraidah Mohd Don* in ***Learner Corpus Studies in Asia and the World***. Shin'Ichiro Ishikawa (Ed). 2013. ISSN 2187-6746	Book Chapter	2013
2.	The overgeneration of *BE+Verb* constructions in the writing of L1-Malay ESL learners in Malaysia *Roslina Abdul Aziz* *Zuraidah Mohd Don* ***Research in Corpus Linguistics Journal (RiCL)*** ISSN 2243-4712	Journal	2014
3.	Tranferable text processing technologies: Assessing performance of a part-of-speech tagger on Malaysian Corpus of Learner English Manuscript sent to: ***Revista De Linguistica (RLA)***	Journal (In progress)	2018
4.	Corpus Consultation in Language Teaching: A Proposed Model In ***Teaching and Learning English in Malaysian Higher Education: Sharing Experience to Improve Practice***. Zuraidah Mohd Don, Siti Zaidah Zainuddin, Tam Shu Sim & Azlin Zaiti Zainal (Eds.) Universiti of Malaya Press	Book Chapter (In progress)	2018

B. List of Presentations

No	Title	Presentation	Year
1.	The *BE* verb omission among advanced L1-Malay ESL learners: What Corpus-based study can reveal. *Roslina Abdul Aziz* *Zuraidah Mohd Don*	International Symposium on Learner Corpus Studies in Asia and the World (LCSAW) 2013	23-24 March 2013
2.	A corpus-based investigation into *BE* overgeneration constructions by L1-Malay ESL learners in the Malaysian Corpus of Learner English (MACLE) *Roslina Abdul Aziz* *Zuraidah Mohd Don*	Asia Pacific Corpus Linguistics Conference 2012 University of Auckland	15-19 Feb 2012